Preface

Transportation sustains economic and social activity and is central to operations research and management science. When operations research emerged as a structured field during World War II, some of the first problems investigated arose from the need to optimize military logistics and transportation activities. After the war ended, the scope of operations research applications broadened but transportation problems always occupied a central place. It is now widely recognized that some of the most successful applications of operations research are encountered in transportation, most significantly in the airline industry where it underlies almost every aspect of strategic, tactical, and operational planning. This success story may be explained by a number of factors, the first being the economic importance of transportation. Also, the complexity and large scale of transportation problems call for powerful analytical techniques, and the high volumes involved imply that substantial savings can often be achieved through the use of optimization. Furthermore, transportation problems are highly structured, making them amenable to the use of efficient solution methods based on network optimization techniques and mathematical programming.

This book contains eleven chapters describing some of the most recent methodological operations research developments in transportation. It is structured around the main transportation modes, and each chapter is written by a group of well-recognized researchers. Because of the major impact of operations research methods in the field of air transportation over the past forty years, it is befitting to open the book with a chapter on airline operations management. While many past publications have focused on airline strategic and tactical planning, Ball, Barnhart, Nemhauser, and Odoni have chosen to address the organization and control of recovery operations in the event of disturbances. This line of research is relatively new and of major importance to the airline industry. The second chapter, by Desaulniers and Hickman, surveys the planning of public transit operations. The problems addressed and the methods employed in transit planning, for example, those arising in network design, passenger assignment, scheduling, and fleet and crew assignment, are often similar to those of the airlines. The railway optimization chapter, by Caprara, Kroon, Monaci, Peeters, and Toth, covers the realm of planning problems encountered in railway planning, with an emphasis on European passenger railways. Again, several of these issues are similar to those observed in other modes, but some problems are specific to the railway industry, such as train platforming, rolling stock circulation, and train unit shunting. The fourth

chapter, by Christiansen, Fagerholt, Nygreen, and Ronen, contains an extensive survey of maritime transportation problems, methods, and applications. Compared with other modes, maritime transportation has received relatively little attention from operations researchers. Yet this field is rapidly expanding with the consolidation of major shipping companies and the development of large container ports.

The next three chapters cover a variety of planning problems arising in vehicle fleet management. The chapter by Powell, Bouzaïene-Ayari, and Simão addresses truck transportation planning in contexts where information processes are dynamic. The focus is on the development of models that capture the flow of information and decisions. The vehicle routing chapter, by Cordeau, Laporte, Savelsbergh, and Vigo, concerns what is arguably the most central problem in distribution management. It surveys several families of vehicle routing problems, including classical models, inventory routing, and stochastic routing. In the transportation on demand chapter, Cordeau, Laporte, Potvin, and Savelsbergh consider the planning of pickup and delivery operations made at the request of users, such as those encountered in courier services, dial-a-ride operations, dial-a-flight systems, and ambulance fleet deployment.

The eighth chapter, by Crainic and Kim, is devoted to intermodal transportation and ties in some planning issues encountered in railway, maritime, and trucking operations. This chapter describes methodologies relevant to the solution of system design and operations planning problems from the perspective of a carrier, or from that of an intermodal transfer facility operator. It also addresses problems encountered at the regional or national level. The next chapter, by Erkut, Tjandra, and Verter, concerns the transportation of hazardous materials and includes a broad description of the issues encountered in this field, as well as methodological contributions on risk assessment, routing and scheduling, and facility location.

The last two chapters of the book cover the area of automobile transportation. Marcotte and Patriksson first survey the broad field of traffic equilibrium. Their chapter contains a rich account of the main equilibrium concepts, as well as subproblems and mathematical algorithms encountered in this area. This chapter provides an informative bibliographical note at the end of each section. Finally, in the last chapter, Papageorgiou, Ben-Akiva, Bottom, Bovy, Hoogendoorn, Hounsell, Kotsialos, and McDonald summarize some of the most important issues and recent developments encountered in ITS and traffic management. These include traffic flow models, route guidance and information systems, as well as urban and highway traffic control.

We are confident that this book will prove useful to researchers, students, and practitioners in transportation, and we hope it will stimulate further research in this rich and fascinating area. We are grateful to Jan Karel Lenstra and George L. Nemhauser who invited us to edit this volume. While the process took longer than we had expected, we found the experience highly rewarding. Our deep thanks go to all authors for the quality of their contri-

butions, to the anonymous referees for their time, effort, and valuable suggestions, and to Gerard Wanrooy of Elsevier for his support.

Cynthia Barnhart
Massachusetts Institute of Technology

Gilbert Laporte
HEC Montréal

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14001-3

Chapter 1

Air Transportation: Irregular Operations and Control

Michael Ball

Decision and Information Technologies Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, USA
E-mail: mball@rhsmith.umd.edu

Cynthia Barnhart

Civil and Environmental Engineering Department and Engineering Systems Division, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
E-mail: cbarnhart@mit.edu

George Nemhauser

Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
E-mail: george.nemhauser@isye.gatech.edu

Amedeo Odoni

Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
E-mail: arodoni@mit.edu

1 Introduction

Commercial aviation operations are supported by what is probably the most complex transportation system and possibly the most complex man-made system in the world. Airports make up the fixed "nodes" on which the system is built. Aircraft represent the very valuable assets that provide the basic transportation service. Passengers demand transportation between a multitude of origins and destinations, and request specific travel dates and times. Crews of pilots and flight attendants operate the aircraft and provide service to passengers. These disparate entities are coordinated through a flight schedule, comprised of flight legs between airport locations. The flight schedule itself defines three other layers of schedules, namely the *aircraft schedule*, the *crew schedule*, and *passenger itineraries*. The aircraft schedule is an assignment of the legs in the flight schedule, with each aircraft assigned to a connected sequence of origin to destination flight legs. When an aircraft carries out a flight between an origin and destination airport, it follows a *flight plan* that defines a sequence of points in the airspace through which it proceeds. The crew schedule is an assignment of the legs in the flight schedule to pilots and flight attendants, ensuring that all crew movements and schedules satisfy collective bargaining

agreements and government regulations. Passenger schedules, which represent the end-customer services, define the familiar itineraries consisting of lists of origin and destination airports together with scheduled arrival and departure times. Typically, pilots and flight attendants have distinct schedules. The itineraries of a specific crew and a specific aircraft may coincide for several flight legs, but they, like passengers and aircraft, often separate at some point during a typical day's operations.

The fact that a single flight leg is a component of several different types of schedules implies that a perturbation in the timing of one leg can have significant "downstream" effects leading to delays on several other legs. This "fragility" is exacerbated by the fact that most of the largest carriers rely heavily on hub-and-spoke network configurations that tightly inter-connect flights to/from many different "spokes" at the network's hubs. Thus, any significant disturbance at a hub, rapidly leads to disruptions of extensive parts of the carrier's schedules. Notable categories of events leading to such disruptions include:

1. *Airline resource shortages* stemming from aircraft mechanical problems, disrupted crews due to sickness, earlier upstream disruptions, longer than scheduled aircraft turn times caused by lack of ground resources to operate the turn, longer than expected passenger embarking and disembarking times, or delayed connecting crews or connecting passengers.
2. *Airport and airspace capacity shortages* due to weather or to excessive traffic. Inclement weather is cited as the source of 75% of airline disruptions in the United States (Dobbyn, 2000).

In 2000, about 30% of the jet-operated flight legs of one major US airline were delayed, and about 3.5% of these flight legs were canceled. Yu et al. (2003) report for another major US airline that, on average, a dozen crews are disrupted every day. The effects of these disruptions are exacerbated when applied to *optimized* airline schedules, for which cost minimization and intensive resource utilization tend to go hand-in-hand. Nonproductive resources, such as idle aircraft and crew on the ground, are costly. Hence, optimized schedules have minimal nonproductive, or *slack*, time between flight legs. In these finely tuned, optimized schedules, delay often propagates with no slack to absorb it, making it very difficult to contain disruptions and to recover from their effects. A mechanical delay affecting a single aircraft can result in delays to passengers and crews assigned to aircraft other than those delayed, due to the interconnectivity of passengers, crews, and aircraft. This network propagation phenomenon explains why weather delays in one geographical area, delaying flights in and out of that area, can result in aircraft, crew, and passenger delays and cancellations in locations far removed from the weather delay. In fact, such *local* delays can impact network operations *globally*.

The significance of the delay propagation effect is illustrated in Figure 1 (reprinted from Beatty et al., 1998). This graphic is based on an analysis of American Airlines passenger and aircraft schedule information. The x-axis

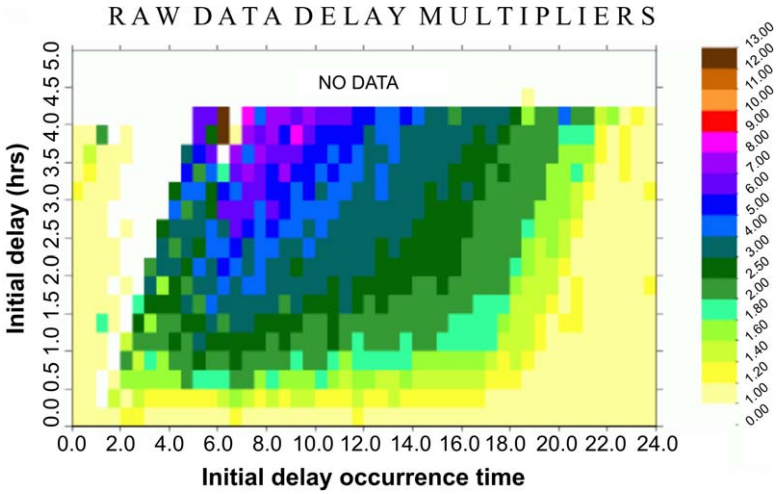


Fig. 1. Estimation of delay propagation multiplier (from Beatty et al., 1998).

tracks time of day from early morning to evening. The y-axis tracks increasing values of an initial flight delay. The color of each box in the x–y plane corresponds to the multiplier that can be applied to an initial delay to estimate the impact of delay propagation. For example, an initial delay of 1.5 hours at 8:00 is colored dark green indicating a delay multiplier of 2.5. This means that an original delay of 1.5 hours on a particular flight induces $2.5 \cdot 1.5 = 3.75$ hours in total flight delay. Note that the delay multiplier increases with the size of the original delay and is greatest during the peak morning periods.

The economic impact of disruptions is great. According to Clarke and Smith (2000), disruption costs of a major US domestic carrier in one year exceeded $440 million in lost revenue, crew overtime pay, and passenger hospitality costs. Moreover, the Air Transport Association (*http://www.airlines.org/econ/files/zzzeco32.htm*) reported that delays cost consumers and airlines about $6.5 billion in 2000. These costs are expected to increase dramatically, with air traffic forecast to double in the next 10–15 years. The MIT Global Airline Industry Program (*http://web.mit.edu/airlines/industry.html*) and Schaefer et al. (2005) indicate that, at current demand levels, each 1% increase in air traffic will bring about a 5% increase in delays.

In this chapter we consider problems related to the management of air traffic and airline operations for the purpose of minimizing the impact and cost of disruptions. The considerable system complexity outlined above makes these problems challenging and has motivated a vibrant and innovative body of research. We start in Section 2 by providing background which is essential to understanding the fundamental issues and motivating the subsequent material. We first review the "physics" and characteristics of airspace system elements and airspace operations in order to explain why capacity constraints are so unpredictable and variable from day to day. Of critical importance are the arrival

and departure capacities of airports, which depend on weather, winds, and the number of active runways and their configuration.

Providers of air traffic control services, such as the Federal Aviation Administration (FAA) and EUROCONTROL, have responsibility for overall airspace management and as such are interested in achieving high levels of system-wide performance. The two broad classes of "tools" at their disposal include restricting schedules and air traffic flow management (ATFM). The former tool, which is treated in Section 3, is strategic in nature. It seeks to control or influence the airline schedule-planning process by ensuring that the resultant schedules do not lead to excessive levels of system congestion and delays. Restricting schedules is particularly challenging in that the competing economic interests of multiple airlines must be balanced. In fact, recent research in this area has been investigating the potential use of market-based mechanisms for this purpose, including auctions and peak-period pricing. The second tool, ATFM, is tactical in nature and is treated in Section 4. ATFM encompasses a broad range of techniques that seek to maximize the performance of the airspace system on any given day of operations, while taking into account a possibly broad range of disruptive events. Many ATFM actions and solutions involve an allocation of decision-making responsibilities between the air traffic control service provider and an airline. This is most notably the case for solutions employing the Collaborative decision making (CDM) paradigm, which has the explicit goal of assigning decision making responsibility to the most appropriate stakeholder in every case (Section 4.4).

A brief Section 5 describes some simulation models that can be useful support tools in understanding and visualizing the impact of certain types of disruptive events on airport, airspace, and airline operations and on air traffic flows, as well as in testing the effectiveness of potential responsive actions.

Sections 6 and 7 address schedule planning and operations problems from the airline perspective. The introductory paragraphs of this section described several factors, which lead to the high complexity of these problems. This complexity is compounded by two additional important considerations:

1. The predominant concern with safety and the significant unionization of crews that, in combination, have led to the *imposition of a very large set of complicated constraints* defining feasibility of flight plans and of flight aircraft and crew schedules.
2. The *size of airline networks and operations*, including, in the United States alone, over 5000 public-use airports serving over 8000 (nongeneral aviation) aircraft transporting approximately 600 million passengers on flights covering more than 5 billion vehicle miles annually (*http://www.bts.gov/publications/pocket_guide_to_transportation/2004/pdf/entire.pdf*).

The aircraft- and crew-scheduling problem, also referred to as the *airline schedule planning* problem, involves designing the flight schedule and assigning aircraft, maintenance operations and crews to the schedule. The typical size of this problem is so large that it is impossible to solve it directly for large

airlines. Instead, airlines partition it into four subproblems, namely: (i) schedule generation; (ii) fleet assignment; (iii) maintenance routing; and (iv) crew scheduling. The subproblems are solved sequentially, with the solutions to the earlier, higher-level subproblems serving as the fixed inputs to subsequent ones. The schedule generation problem is to determine the flight legs, with specified departure times, comprising the flight schedule. These legs, which define the origin–destination markets served and the frequency and timing of service, have significant effects on the profitability of airlines. Given the flight schedule, the fleet assignment problem is to find the profit maximizing assignment of aircraft types to flight legs in the schedule. Where possible, the goal is to match as closely as possible seat capacity with passenger demand for each flight leg. With the fleeted flight schedule and the size and composition of the airline's fleet as input, the maintenance routing problem is to find for each aircraft, a set of *maintenance-feasible rotations*, or routes that begin and end at the same place and satisfy government- and airline-mandated maintenance requirements. Finally, given all the schedule design and aircraft assignment decisions, the crew scheduling problem is to find the cost minimizing assignment of cockpit and cabin crews to flights. Crew costs, second only to fuel costs, represent a significant operating expense. A detailed description of the airline schedule planning problem is provided in Barnhart et al. (2003a).

From the airline perspective, the focus of this chapter is motivated by the fact that, despite advances in aircraft and crew schedule planning, *optimized* plans are rarely, if ever, executed. Thus, we shall not provide broad coverage of airline schedule planning but rather focus on the topics that address the development of schedules and operating practices and policies that provide operational robustness. In Section 6, we cover the theme of optimizing airline schedule recovery. The associated tools are designed for use in a near real-time mode to adjust operations in response to a variety of disruptions. In Section 7, we address, by contrast, the more strategic topic of developing schedules that provide operational robustness. This more recent area of study builds upon the long-standing and well-known body of research on aircraft and crew scheduling described in the previous paragraph. Finally, in Section 8, we conclude with a very general assessment of the state of research and implementation in this subject area.

2 Flow constraints in the infrastructure of commercial aviation

The airspace systems of developed nations and regions consist of a set of often extremely expensive and scarce nodes, the airports, and of air traffic management (ATM) systems that provide aircraft and pilots with the means needed to fly safely and expeditiously from airport to airport. The essential components of ATM systems are a skilled workforce of human air traffic controllers; organization of the airspace around airports ("terminal airspace") and

between airports ("en-route airspace") into a complex network of *airways*, *waypoints*, and *sectors* of responsibility; procedures and regulations according to which the ATM system operates; automation systems, such as computers, displays, and decision support software; and systems for carrying out the functions of communications, navigation, and surveillance (CNS) which are critical to ATM. Any flow constraints encountered during a flight may result from an obvious cause (e.g., the closing down of a runway) or from a set of complex interactions involving failure or inadequacy of several of the components of the ATM system.

A *controlled flight* is one for which an approved *flight plan* has been filed with the air traffic management (ATM) system. Airline and general aviation operators prepare and file flight plans usually based on criteria that consider each flight in isolation. Air carriers typically employ sophisticated software, including advanced route optimization programs, for this purpose. Far from being just a "shortest path" problem, the selection of an optimal route for a flight typically involves a combination of criteria, such as minimum time, minimum fuel consumption, and best ride conditions for the passengers. Midkiff et al. (2004) provide a thorough description of air carrier flight planning. By accepting a flight plan, the ATM system agrees to take responsibility for the safe separation of that aircraft from all other controlled aircraft in the airspace and to provide many other types of assistance toward the goal of completing the flight safely and expeditiously. Practically all airline flights and a large number of general aviation flights are controlled. The focus of this entire chapter is on flow constraints that such flights often face and on how ATM service providers, airport operators, and airlines attempt to deal with them. Figure 2 illustrates schematically the fact that such constraints or "bottlenecks" occur when flights are departing from and arriving at airports and when they seek to access certain parts of the airspace.

This section will seek to review the "physics" of the constraints, with emphasis on explaining the causes of two of their most distinctive characteristics – variability and unpredictability. It is these characteristics that make the constraints so difficult to deal with in practice, as well as so interesting for many researchers. Emphasis will be given to the specific topic of capacity-related flow

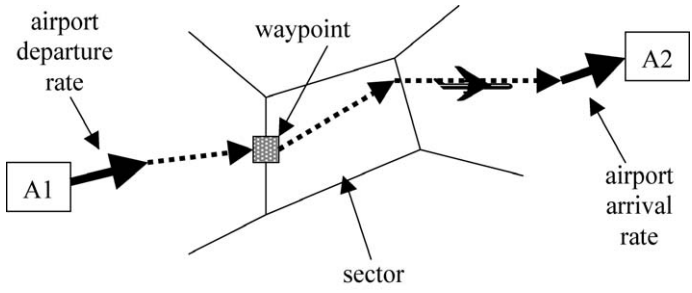


Fig. 2. Airspace flow constraints.

constraints at major commercial airports, due to the enormous practical significance and cost consequences of these constraints.

2.1 The "physics" of airport capacity

Airports consist of several subsystems, such as runways, taxiways, apron stands, passenger and cargo terminals, and ground access complexes, each with its own capacity limitations. At major airports, the capacity of the system of runways is the most restricting element in the great majority of cases. This is particularly true from a long-run perspective. While it is usually possible – albeit occasionally very expensive – to increase the capacity of the other airport elements through an array of capital investments, new runways, and associated taxiways require great expanses of land and have environmental and other impacts that necessitate long and complicated approval processes, often taking a couple of decades or even longer, with uncertain outcomes.

The capacity of runway systems is also the principal cause, by far, of the most extreme instances of delays that lead to widespread schedule disruptions, flight cancellations, and missed flight connections. There certainly have been instances when taxiway system congestion or unavailability of gates and aircraft parking spaces have become constraints at airports, but these are more predictable and stable. The associated constraints can typically be taken into consideration in an ad hoc way during long-range planning or in the daily development of ATFM plans (Section 4). By contrast, the capacity of the runway system can vary greatly from day to day and the changes are difficult to predict even a few hours in advance. This may lead to an unstable operating environment for air carriers: on days when an airport operates at its nominal, good-weather capacity, flights will typically operate on time, with the exception of possible delays due to "upstream" events; but, with the same level of demand at the same airport, schedule reliability may easily fall apart on days when weather conditions are less than ideal.

The capacity of a runway, or of a set of simultaneously active runways at an airport, is defined as the expected number of movements (landings and takeoffs) that can be performed per unit of time in the presence of continuous demand and without violating air traffic control (ATC) separation requirements. This is often referred to as the *maximum throughput capacity*, C. Note that this definition recognizes that the actual number, N, of movements that can be performed per unit of time is a random variable. The capacity C is simply defined as being equal to $E[N]$, the expected value of N. The unit of time used most often is one hour.

To understand better the multiple causes of capacity variability, especially its strong dependence on weather and wind conditions, it is necessary to look at the "physics" of the capacity of runway systems. It is convenient to consider first the case of a single runway and then (Section 2.1.4) the case of a system of several runways.

2.1.1 Factors affecting the capacity of a single runway

The capacity of a single runway depends on many factors, the most important of which are:

1. The mix of aircraft classes using the airport.
2. The separation requirements imposed by the ATM system.
3. The type (high speed or conventional) and location of exits from the runway.
4. The mix of movements on each runway (arrivals only, departures only, or mixed) and the sequencing of the movements.
5. Weather conditions, namely visibility, cloud ceiling, and precipitation.
6. The technological state and overall performance of the ATM system.

The impacts of 1–4 are summarized below, while 5 and 6 are discussed in the more general context of multi-runway systems in the next subsection.

Mix of aircraft. The FAA and other Civil Aviation Authorities around the world classify aircraft into a small number of classes for terminal area ATC purposes. For example, the FAA defines four classes, based on maximum take-off weight (MTOW): "Heavy" (H), "Large" (L), the Boeing 757 (a class by itself), and "Small" (S). Most other Civil Aviation Authorities have adopted the same or very similar classifications. Roughly speaking, the H class includes all wide-body jets, and the L class practically all narrow-body commercial jets – including many of the larger, new generation, regional jets – as well as some of the larger commercial turbo-props. Most general aviation airplanes, including most types of private jets, as well as the smaller commercial turboprops and regional jets with about 35 seats or fewer comprise the S class. The *aircraft mix* indicates the composition of the aircraft fleet that is using any particular runway (e.g., 20% S, 60% L, 5% B757, and 15% H).

Separation requirements and high-speed exits. The single most important factor in determining runway capacity is the separation requirements, which impose safety-related separations between aircraft that limit the service rate of the runway, i.e., its maximum throughput capacity. For every possible pair of aircraft using the same runway consecutively, the FAA and other Civil Aviation Organizations specify a set of separation requirements in units of distance or of time. These requirements depend on the classes to which the two aircraft belong and on the types of operation involved: arrival followed by arrival, "A–A", arrival followed by departure, "A–D", etc. Table 1 shows the separation requirements that currently apply at most of the busiest airports in the United States for the case in which a runway is used *only for arrivals* under *instrument flight rules* (IFR). Pairs of consecutive landing aircraft must maintain a separation equal to or greater than the distances indicated in Table 1 throughout their final approach to the runway, with the exception of the cases marked with an asterisk, where the required separation must exist at the instant when the leading aircraft reaches the runway. The 4, 5, and 6 nautical mile separations

Table 1.
FAA IFR separation requirements in nautical miles (nmi) for "an arrival followed by an arrival". Asterisks indicate separations that apply when the leading aircraft is at the threshold of the runway.

Leading aircraft	Trailing aircraft		
	H	L + B757	S
H	4	5	6*
B757	4	4	5*
L	2.5	2.5	4*
S	2.5	2.5	2.5

shown in Table 1 are intended to protect the lighter trailing aircraft in the pair from the hazards posed by the wake vortices generated by the heavier leading aircraft. These are therefore often referred to as "wake vortex separations". In addition to the "airborne separation" requirements of Table 1, a further restriction is applied: the trailing aircraft of any pair cannot touch down on the runway before the leading aircraft is clear of the runway. In other words, the runway can be occupied by only one arriving aircraft at any time.

The more restrictive of the two requirements – "airborne separation" and "single occupancy" – is the one that applies for each pair of aircraft. When arrivals take place in instrument meteorological conditions (IMC), "airborne separation" is almost always the most restrictive. However, "single occupancy" may become the constraint when visual airborne separations on final approach are allowed (instead of the distance requirements of Table 1), as is often done in the United States under visual meteorological conditions (VMC). In this case, high-speed runway exits and well-placed runway exits (the third of the factors identified above), which reduce runway occupancy times for arriving aircraft, can be helpful in increasing runway capacity. High-speed exits can also be useful when the runway is used for both arrivals and departures: if landing aircraft can exit a runway quickly, air traffic controllers may be able to "release" a following takeoff sooner.

Mix of movements. Separation requirements, analogous to those in Table 1, are also specified for the other three combinations of consecutive operations, A–D, D–A, and D–D. Because the separation requirements for each combination are different – see, e.g., Chapter 10 of de Neufville and Odoni (2003) for details – the capacity of a runway during any given time period depends on the mix of arrivals and departures during that period, as well as on how exactly arrivals and departures are sequenced on the runway. This also suggests that there is an important tradeoff between the maximum arrival and departure rates that an airport can achieve.

2.1.2 Capacity envelope and its computation

The runway *capacity envelope* (Figure 3) is convenient for displaying the arrival and departure capacities and associated tradeoffs. The capacity envelope of a single runway is typically approximated by a piecewise linear boundary that connects four points (Gilbo, 1993). Points 1 and 4 indicate the capacity of the runway, when it is used only for arrivals and only for departures, respectively. Point 2 is known as the "free-departures" point because it has the same capacity for arrivals as Point 1 and a departures capacity equal to the number of departures that can be inserted into the arrivals stream without increasing the separations between successive arrivals – and, thus, without reducing the number of arrivals from what can be achieved in the all-arrivals case. These "free" departures are obtained by exploiting large inter-arrival gaps such as the ones that arise between a "H-followed-by-S" pair of landing aircraft. Point 3 can be attained, in principle, by alternating arrivals and departures, i.e., by performing an equal number of departures and arrivals through an A–D–A–D–A–· · · sequence. This sequencing strategy can be implemented by "stretching", when necessary, inter-arrival (inter-departure) gaps by an amount of time just sufficient to insert a departure (arrival) between two successive arrivals (departures). Because it is difficult for air traffic controllers to sustain this type of operation for extended periods of time, Point 3 can be viewed as somewhat theoretical. However, it provides a useful upper limit on the total achievable capacity (landings plus takeoffs) when arrivals and departures share a runway in roughly equal numbers.

Several mathematical models have been developed over the years for computing the capacity of a single runway under different sets of conditions, beginning with Blumstein's (1959) classical model of a single runway used for arrivals only. The models have become increasingly sophisticated

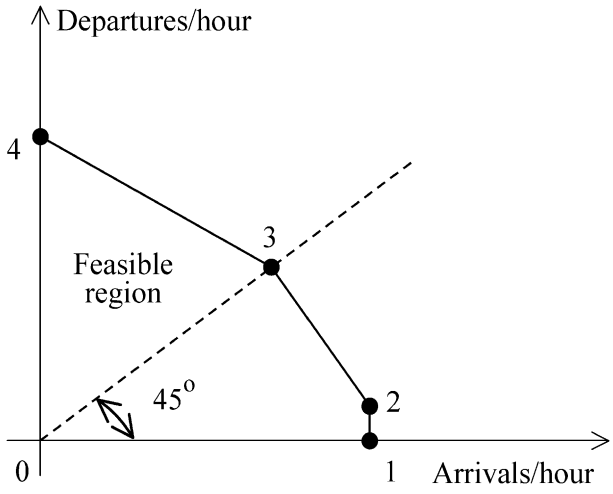


Fig. 3. A capacity envelope for a single runway.

over the years and include treatment of some of the input parameters as random variables. Barnhart et al. (2003a) provide a literature review. The most recent of these models (Long et al., 1999; Stamatopoulos et al., 2004; EUROCONTROL, 2001) incorporate most of the best features of earlier models and generate capacity envelopes, such as the one in Figure 3.

2.1.3 The sequencing problem

As a result of the airborne separation requirements shown in Table 1, certain aircraft pairs require longer separation distances than others and thus the total time needed for the landing of any set of aircraft on a runway depends on the sequencing of the aircraft. For example, the "H followed by S" sequence will consume much more time than "S followed by H". Given a number n of aircraft, all waiting to land on a runway, the problem of "determining the sequence of landings such as to minimize the time when the last aircraft lands" is a Hamiltonian path problem with n points (Psaraftis, 1980; Venkatakrishnan et al., 1993). This is a problem entirely analogous to several well-known job-sequencing problems in manufacturing.

However, the Hamiltonian path approach addresses only a static version of a problem. In truth, the problem of sequencing aircraft on a runway is dynamic: over time, the pool of aircraft available to land changes, as some aircraft reach the runway while new aircraft join the arrivals queue. Moreover, minimizing the "latest landing time" (or maximizing "throughput") should not necessarily be the objective of optimal sequencing. Many alternative objective functions, such as minimizing the average waiting time per passenger, are just as reasonable. A further complication is that the very idea of "sequencing" runs counter to the traditional adherence of ATM systems to a first-come, first-served (FCFS) discipline, which is perceived by most as "fair" (see also Section 4).

These observations have motivated a great deal of research on the *runway sequencing problem* with the objective of increasing operating efficiency while ensuring that all airport users are treated equitably. Dear and Sherif (1991) developed the concept of *constrained position shifting* (CPS), i.e., of a limit in the number of positions by which an aircraft can deviate from its FCFS position in a queue. For instance, an aircraft in the 16th position in an FCFS queue, would have to land in one of the positions 14–18, if the specified *maximum position shift* (MPS) is 2. Through many numerical examples and for several reasonable objective functions, Dear showed that, by setting MPS to a small number, such as 2 or 3, one can obtain most (e.g., 60–80%) of the potential benefits offered by unconstrained optimal sequences and, at the same time, ensure reasonable fairness in accessing runways. Several researchers (e.g., Psaraftis, 1980; Venkatakrishnan et al., 1993; Beasley et al., 2001) have investigated a number of increasingly complex and realistic versions of the sequencing problem. Two advanced terminal airspace automation systems, CTAS and COMPAS, that have been implemented in the US and in Germany, respectively, incorporate sequencing algorithms based on CPS (Erzberger, 1995).

Gilbo (1993), Gilbo and Howard (2000), and Hall (1999) have gone beyond the sequencing of arrivals only, by considering how available capacity can best be allocated in a dynamic way between landings and take-offs to account for the distinct peaking patterns in the arrival and departure streams at airports over the course of a day. They propose the application of optimization algorithms that use capacity envelopes (Figure 3) within the context of ATFM to achieve an optimal trade-off between arrival and departure rates and, by implication, between delays to arrivals and to departures.

2.1.4 Factors affecting the capacity of multi-runway systems

Most (but certainly not all) major airports typically operate with two or more simultaneously active runways. The term *runway configuration* refers to any set of one or more runways, which can be active simultaneously at an airport. Multi-runway airports may employ more than ten different runway configurations. Which one they will operate on at any given time will depend on weather and wind conditions, on demand levels at the time and, possibly, on noise considerations, as will be explained below.

The six factors listed in Section 2.1.1 clearly continue to affect the capacity of each individual active runway in multi-runway cases. In addition, at least four other factors may now play a major role:

7. The interactions between operations on different runways, as determined by the geometric layout of the runway system and other considerations.
8. The allocation of aircraft classes and types of operations (arrivals, departures, mixed) among the active runways.
9. The direction and strength of winds.
10. Noise-related and other environmental considerations and constraints.

Interactions between operations on different runways. The influence of the geometric layout on the interactions among runways can most simply be illustrated by looking at situations involving two parallel runways. Depending on the distance between their centerlines, operations on the two runways may have to be coordinated all the time, or may be dependent in some cases and independent in others, or may be completely independent. For example, in the United States, two parallel runways separated by less than 2500 ft (762 m) must be operated with essentially single runway separations in instrument meteorological conditions (IMC). This means, for instance, that, if two arriving aircraft are landing, one on the left runway and the other on the right runway, they are subject to the same set of airborne separation requirements as shown in Table 1 for a single runway. At the opposite extreme, if the centerlines are separated by more than 4300 ft (1310 m) the two runways may be operated independently and can accept simultaneous parallel approaches. (With special instrumentation, the FAA will consider authorizing independent parallel approaches with centerline separations as small as 3000 ft (915 m).) Finally, for

intermediate cases, contemporaneous arrivals on the two parallel runways are treated as "dependent", i.e., must be coordinated, but an arrival on one of the runways and a contemporaneous departure from the other can be handled independently. It follows that the combined capacity of the two runways will be highest in the case of independent operations, intermediate in the "partially dependent" operations case, and lowest when every pair of operations on the two runways must be coordinated. In a similar way, the combined capacity of pairs of intersecting runways or of runways that do not intersect physically but intersect at the projections of their centerlines depends on many geometry-related parameters such as the location of the intersection or of the projected intersection; the direction of operations on the two runways; the types of operations and the mix of aircraft; and obviously, the separation requirements for the particular geometric configuration at hand. Systems of three or more active, nonparallel runways typically involve even more complex interactions.

Allocation of aircraft and operations. With more than one active runway, there is some opportunity to "optimize" operations by judiciously assigning operations and/or aircraft classes to different runways. For example, in the case of intermediately spaced parallel runways (centerline separations of 2500–4300 ft in the United States) it may be advisable to use one runway primarily for arrivals and the other primarily for departures. Since, in this case, arrivals on one runway can operate independently of departures on the other, this allocation strategy minimizes interactions between runways and reduces controller workload. Similarly, when two or more runways are used for arrivals, air traffic managers often try to assign relatively homogeneous mixes of aircraft to each of the runways, e.g., keep the "Small" aircraft on a separate runway from the "Heavy" and "Large", to the extent possible. In this way, air traffic controllers can avoid the extensive use of the 5 and 6 nautical mile wake-vortex separations that are required when a Small aircraft is landing behind a Large or a Heavy (Table 1).

Weather-related factors. It is easy to infer from what has been said so far that weather-related factors (numbers 5 and 9 in our list) are critical in determining the variability of the capacity of any system of runways. First, for individual runways, the actual separations between consecutive operations are strongly influenced by visibility, cloud ceiling, and precipitation. This is especially true in the United States where, in good weather, pilots are usually requested to maintain visual separations during the final approach phase from the aircraft landing ahead of them. This practice results in somewhat closer spacing of landing aircraft than suggested by the IFR separations of Table 1. It also means smaller deviations from the required minima, as pilots can adjust spacing as they approach the runway. This second effect is also present at airports where the practice of "visual separations on final" in VMC has not yet been adopted. The overall effect is that, with the same aircraft mix and the same nominal

(IFR) separation requirements, the capacity of individual runways in VMC is typically greater than in IMC, occasionally by a significant margin.

Second, when it comes to multi-runway configurations, good visibility conditions have a similar effect. For example, in the case of two parallel runways, the FAA generally authorizes simultaneous parallel approaches in IMC, when the separation between runway centerlines is 4300 ft or more, as noted earlier. But in VMC, simultaneous parallel approaches can be performed to parallel runways separated by only 1200 ft (366 m) when Heavy aircraft are involved and by only 700 ft (214 m) when they are not. As a consequence, San Francisco International (SFO), one of the most delay-prone airports in the world, has an arrival capacity of about 54 per hour in VMC, when simultaneous parallel approaches are performed to a pair of closely-spaced parallel runways, and of only 34 per hour in IMC, when the same two runways are operated essentially as a single runway, as described earlier. The impact of VMC is similar when it comes to operations on a pair of intersecting runways, as illustrated by New York's LaGuardia Airport (LGA). In VMC, LGA has a nominal capacity of 81 movements per hour and, with the same two runways, a capacity of 63 (or 22% less) in IMC.

Wind direction and wind strength are just as critical in determining which runways will be active at a multi-runway airport at any given time. First, landings and takeoffs are conducted into the wind – the maximum allowable tailwind is generally of the order of 5 knots. Thus, the direction of the wind determines the direction in which the active runways are used. Equally important, there are limits on the strength of the *crosswinds* that aircraft can tolerate on landing and on takeoff. For any runway, the crosswind is the component of the wind vector whose direction is perpendicular to the direction of the runway. The crosswind tolerance limits vary according to type of aircraft and to the state of the runway's surface (dry or wet, slippery due to icy spots, etc.). Thus, airports are often forced by crosswinds and tailwinds to utilize configurations that offer reduced capacity. For example, with strong westerly winds, Boston Logan (BOS) is forced to operate with two main runways even in VMC, instead of the customary three. This reduces capacity by about 30 movements per hour from the VMC norm!

State and performance of the ATM system. An obvious underlying premise to all of the above is that a high-quality ATM system with well-trained and experienced personnel is a prerequisite for achieving high runway capacities. To use a simple example, tight separations between successive aircraft on final approach (i.e., separations which are as close as possible to the minimum required in each case) cannot be achieved unless (a) accurate and well-displayed information is available to air traffic controllers regarding the positions of the leading and trailing aircraft, and (b) the controllers themselves are skilled in the task of spacing aircraft accurately during final approach. Major differences exist in this respect between ATM systems in different countries.

Environmental considerations. Finally, runway usage and, by extension, airport capacity at some major airports may be strongly affected by noise-mitigation and other measures motivated by environmental considerations. In the daily course of airport operations, noise is one of the principal criteria used by air traffic controllers to decide which one among several usable alternative runway configurations to activate. (A choice among two or more alternative configurations may exist whenever weather and wind conditions are sufficiently favorable.) Environmental considerations act, in general, as a constraint on airport capacity since they tend to reduce the frequency with which certain high-capacity configurations may be used.

The capacity envelope for multi-runway systems and its computation. The complexity of computing the capacity envelopes of multi-runway airports depends on the complexity of the geometric layout of the runway system and the extent to which operations on different runways are interdependent. The simplest cases, involving two parallel or intersecting runways, can still be addressed through analytical models, because they are reasonably straightforward extensions of single-runway models (Stamatopoulos et al., 2004). Analytical models also provide good approximate estimates of true capacity in cases involving three or more active runways, as long as the runway configurations can be "decomposed" into semi-independent parts, each consisting of one or two runways. This is possible at the majority of existing major airports and at practically every secondary airport.

When such decomposition is not possible or when a highly detailed representation of runway and taxiway operations is necessary, simulation models can be used. General-purpose simulation models of airside operations first became viable in the early 1980s and have been vested with increasingly sophisticated features since then. Two models currently dominate this field internationally: SIMMOD and the Total Airport and Airspace Modeler (TAAM). A report by Odoni et al. (1997) contains detailed reviews of these and several other airport and airspace simulation models and assesses the strengths and weaknesses of each. At their current state of development and with adequate time and personnel resources, they can be powerful tools not only in estimating the capacity of runway systems, but also in studying detailed airside design issues, such as figuring out the best way to remove an airside bottleneck or estimating the amount by which the capacity of an airport is reduced due to the crossing of active runways by taxiing aircraft. However, these simulation models still involve considerable expense, as well as require significant time and effort and, most importantly, expert users.

2.1.5 The variability and unpredictability of the capacity of runway systems

The variability of the capacity of runway systems at major airports can now be easily explained with reference to the previous discussion. There are three main causes of drastic reductions in airport capacity. Two are related to weather: severe events, like thunderstorms or snowstorms; and more routine

weather events, like fog or very strong winds. The third major cause is technical or infrastructure problems, such as air traffic control equipment outage or the temporary loss of one or more runways, due to an incident or accident or to maintenance work. For this last case, it should be noted that major airports schedule runway maintenance carefully, so as to minimize impact on airport traffic.

Thunderstorms and snowstorms are events that pose hazards to aviation. Thus, they impede severely the flow of air traffic into and out of airports and through major portions of affected airspace. They carry the potential for even shutting down airports completely for several hours and, occasionally, for a few days at a time in the case of snowstorms. The more routine events can be much more frequent, such as heavy fog at the San Francisco, Milan, and Amsterdam airports or strong winds in Boston. These typically cause a severe reduction of capacity, from the best levels achievable in VMC to levels associated either with IMC or with nonavailability of some runways due to winds. The FAA in a 2001 study compared the maximum throughput capacities of the 31 busiest commercial airports in the United States under optimum weather conditions, with the capacity of the most frequently used configuration in IMC (FAA, 2001). The study found that, on average, the capacity was reduced by 22% in the latter case, with 8 of the 31 airports experiencing a capacity reduction of 30% or more! Note that other, less frequently used IMC configurations at these airports often have even lower capacities.

The overall effect of weather on an airport's capacity can be summarized conveniently through the *capacity coverage chart* (CCC), which is essentially a plot of the probability distribution of available capacity over an extended period of time such as a year. An example for Boston's Logan International Airport (BOS) is shown in Figure 4. (The CCC is somewhat simplified to indicate only five principal levels of capacity.) It indicates that the capacity varies

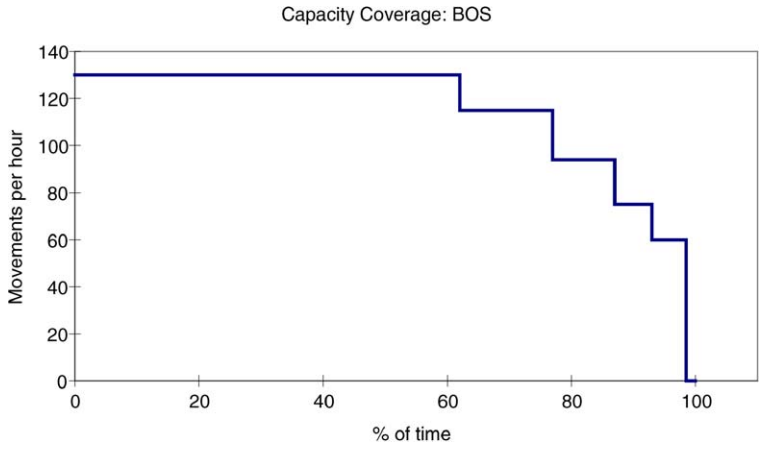


Fig. 4. The capacity coverage chart for Boston Logan International Airport.

from a high capacity of 115 movements or more per hour, available for about 77% of the time – the leftmost two levels of capacity – and associated with the most favorable VMC, to a low of about 55–60 movements per hour for about 6% of the time and associated with low IMC. (The airport also has capacity of zero, meaning it is closed down due to weather conditions, about 1.5% of the time.) One of the two intermediate levels of capacity (third from left) of about 94 movements per hour is associated with the presence of strong westerly winds in VMC. As mentioned in the previous section, these force the airport to operate with only two active main runways. To prepare the CCC, it is necessary to examine historical hourly weather records (visibility, cloud ceiling, precipitation, winds) for a long period of time (e.g., five years) and identify the capacity available at each of these hours.

The CCC is drawn under the simplifying assumptions that (a) the mix of arrivals and departures is 50% and 50% and (b) the airport is operated at all times with the highest-capacity configuration that can be used under the prevailing weather conditions. While neither of these assumptions is exactly true in practice, the CCC nonetheless provides a good indication of the overall availability of capacity during a year, as well as of the variability of this capacity. Obviously a CCC that stays level for the overwhelming majority of time – as one might expect to find at airports that enjoy consistently good weather – implies a more predictable operating environment than the "uneven" CCCs of BOS, SFO, LGA, and other airports where weather is highly variable.

Although the associated technology is improving, meteorological forecasts still have not attained the level of accuracy and detail needed to eliminate uncertainty from predictions of airport and airspace weather, even for a time-horizon as short as one or two hours. When it comes to impact on the operations of any specific airport, the challenge is twofold: predicting the severity of an anticipated weather event at a quite microscopic level; and equally important, determining narrow windows for the forecast starting and ending times of the event. For example, a few hundred feet of difference in the cloud ceiling or the presence or absence of "corridors" for the safe conduct of approaches and departures in convective weather may make a great difference in the amount of capacity available at an airport. Similarly, over- or under-predicting by just one hour the ending time of a thunderstorm may have major implications on a Ground delay program (see Section 4) and, as a result, on the costs and disruption caused by the associated delays and flight cancellations.

2.2 The capacity of airspace sectors

Within the airspace itself, safety concerns and the need to separate aircraft leads to yet another set of constraints. The most prevalent of these is associated with a sector. A sector is a volume of airspace for which a single air traffic control team of one or two individuals has responsibility. The principal constraint on the number of aircraft that can safely occupy a sector simultaneously is controller workload. Both in North American and in European airspace, it is

generally accepted that this number should not exceed the 8–15 range, depending on a number of factors. This limitation, in turn, translates to typical upper limits of the order of 15–20 on the number of aircraft that can be scheduled to traverse a sector during a 15-minute time interval in US en-route airspace. This capacity may be reduced significantly in the presence of severe weather.

Because of its heavy dependence on controller workload, it is difficult to compute the capacity of a sector, in terms of either the number of simultaneously present aircraft or the number of aircraft traversing the sector per unit of time (Wyndemere, 1996; Sridhar et al., 1998). Numerous factors affect the complexity of the controller's task. Hinston et al. (2001) classify these into three major categories:

(a) *Airspace factors*: sector dimensions (physical size and shape, area that the controller must effectively oversee); spatial distribution of airways and of navigational aids within the sector; number and location of standard ingress and egress points for the sector; configuration of traffic flows (number and orientation relative to the shape of the sector, complexity of aircraft trajectories, crossing points and/or merging points of the flows); and complexity of required coordination with controllers of neighboring sectors (e.g., for "hand-offs" of aircraft from one sector to the next).

(b) *Traffic factors*: number and spatial density of aircraft; range of aircraft performance (homogeneous traffic vs. many different types of aircraft with diverse performance characteristics); complexity of resolving aircraft conflicts (which depends on many variables); sector transit time.

(c) *Operational constraints*: restrictions on available airspace, e.g., due to the presence of convective weather or of special use airspace; limitations of communications systems; and procedural flow restrictions at certain waypoints (see Section 4.1.2) or noise abatement procedures in place.

Several attempts have been made in recent years to develop quantitative relationships between some of these factors and controller workload – see, e.g., Manning et al. (2002). To deal with this complexity and handle a large number of aircraft, controllers attempt to introduce a "structure" to the traffic patterns they handle. Examples include (Hinston et al., 2001): spatial standardization of the flows of aircraft within sectors along specific paths; consideration of aircraft in groups, with members of each group linked by common attributes; and concentration around a few "critical points" of the location of potential aircraft encounters or of other occurrences requiring controller intervention.

In order to effect such structuring, flow may usually be directed through a few *waypoints* or *fixes*, which have an associated maximum flow rate. These maximum flow rates can be derived from minimum separation standards or alternatively the maximum rates can be specified by the ATM system itself in order to limit the amount of flow passing downstream. In this latter case, the maximum flow rate can be viewed as a control variable.

Finally, it is noted that airspace constraints typically have a less severe effect on airline operations than airport constraints. This is especially true in the more flexible ATM environment of the United States. The reason, quite

simply, is that a capacity-constrained en-route sector can often be bypassed at limited cost by selecting an alternative route, whereas a flight has no choice but to eventually end up at its destination airport.

3 Restricting schedules

3.1 Options and current practice on airport scheduling

With the background of Section 2, we can now proceed to review how the two principal types of airspace system stakeholders attempt to deal with unpredictable and variable capacity constraints in daily operations. This and the next section will discuss strategies associated primarily with ATM service providers (Civil Aviation Authorities and other national and international organizations) while Sections 6 and 7 will present the strategic and tactical options available to the airlines.

To ATM service providers (e.g., the FAA) two approaches are essentially available. One, *restricting schedules* (RS), is of a static and "pro-active" nature as it places, in advance, limits on the maximum number of aircraft movements that can be scheduled during a unit of time at an airport or other airspace element. The second, *ATFM*, is dynamic and reactive: its goal is to prevent airport and airspace overloading by adjusting in "real time" the flows of aircraft on a national or a regional basis in response to actual conditions. In essence, the focus of RS is on controlling the number of *scheduled* operations through airspace elements and of ATFM on controlling the number of *actual* operations through these elements, *given* a schedule. This section reviews briefly the RS approach, while the next deals more extensively with ATFM.

A far more frequently used term for RS, especially among aviation policymakers, is "*demand management*". It refers to any set of administrative and/or economic policies and regulations aimed at *constraining the demand* for access to airspace elements during certain times when congestion would otherwise be experienced. This term is avoided here, because it may cause confusion with a major aspect of ATFM, which is also concerned with "managing demand" in a dynamic way in order to match it with available capacity.

RS is not used currently in the United States, with the exception of four airports (New York LaGuardia and Kennedy, Chicago O'Hare and Washington Reagan) where limits on the number of movements that can be scheduled per hour – the so-called "high-density rules" (HDR) – have existed since 1968. The HDR will be phased out by 2007 according to the so-called AIR-21 legislation of 2000 and, in fact, in some cases, e.g., Chicago, the restrictions have already been relaxed. However, RS is widely practiced outside the US: about 140 of the world's busiest airports are "fully coordinated", meaning that they place strict limits on the number of movements that the airlines can schedule there. These airports serve the great majority of air travelers outside the US every year.

The concept underlying RS is the *declared capacity*, i.e., a declared limit on the maximum number of air traffic movements that can be scheduled at an airport per unit of time. At a few airports, separate limits are specified for the number of landings, the number of takeoffs and the total number of movements. The typical unit of time is one hour, but some airports use finer subdivisions of time. At a few airports, the declared capacity may also vary by time of day, e.g., more departures than arrivals may be allowed during certain hours and vice versa for other hours.

The declared capacity is determined by the capacity of the most restricting element of the airport – the so-called "bottleneck element". In most cases, this is the runway system of the airport. However, the bottleneck element can also be the passenger terminal, or the apron area, or some other part of the airport. Even in such cases, the computed limit (e.g, the number of passengers that can be scheduled per hour at the passenger terminal) is converted to a declared limit on the number of air traffic movements.

The RS approach – and the concept of declared capacity – can be extended to air traffic control sectors. The declared capacity in this environment is primarily determined by workload considerations, as noted in Section 2.2, taking into consideration traffic patterns, traffic mix, route configuration, etc. EUROCONTROL, the agency that coordinates air traffic management systems over Europe, in effect uses such capacity figures for en-route sectors in its six-month advance planning of traffic loads in European airspace.

3.2 Critical issues regarding restricting schedules

If traffic volumes at airports and airspace sectors are restricted to levels that can be handled comfortably all the time, the RS approach can clearly be effective in reducing major delays and important schedule disruptions. However, the approach is also characterized by several fundamental problems, three of which are described here. First, implicit in the approach is the need to make a tradeoff between delays, on the one hand, and resource utilization, on the other, on the basis of only very aggregate information. Consider, for example, the case of Boston's Logan International Airport (BOS). As suggested by Figure 4, the maximum achievable *arrival* rate at BOS under favorable weather conditions (visibility, cloud ceiling, winds, precipitation) is around 60 per hour. Such conditions prevail about 77% of the time. During the other 22%, the maximum arrival rate is lower – and can be as low as 30 per hour for about 6% of the time. Were BOS to declare an arrival capacity of 60 (and assuming that the airlines and general aviation operators actually scheduled that many movements), delays at BOS would reach high to unacceptable levels during about 22% of the peak traffic hours over any extended span of time, such as a year. On the other hand, declaring an arrival capacity of 30 would practically ensure the absence of serious delays, but would result in gross underutilization of the airport's resources most of the time. In general, one should note that any choice of the value of the declared capacity must be made on the basis of very

aggregate statistical information: when an airport "declares a capacity" for the next six months (as is currently done under the process organized by IATA, see below), all it has to go on is historical statistics about weather conditions at the airport and the resulting available capacity. By contrast, as described in Section 4, the ATFM approach uses real-time capacity forecasts with a maximum time-horizon of about 12 hours. In this light, it is not surprising that no single, internationally accepted methodology exists today for determining and setting the declared capacity of airports and airspace sectors. Practices vary greatly from country to country and, in some cases, from airport to airport within the same country. Some major international airports in Europe and in Asia clearly opt for allowing for a considerable margin of "comfort" by declaring very "conservative" capacities, i.e., capacities near the low end of the available range. This results in wasting significant amounts of extremely valuable capacity.

The second problem with RS is the way it is currently practiced. The declared capacities of the major airports, as updated at six-month intervals, are communicated to international aviation authorities and to the airlines and serve as the basis for scheduling airline operations at busy airports during the international Schedule Coordination Conferences (SCC) that are organized by the International Air Transport Association (IATA) and take place in November and June every year. During the SCC, a "schedule coordinator" allocates the available capacity ("slots") among the airlines that have requested access to each fully coordinated airport. If the number of available slots is insufficient to satisfy demand, then some requests are simply denied. The criteria used to allocate slots are described in detail in IATA (2000) and discussed in de Neufville and Odoni (2003). For our purposes, it is sufficient to note that the dominant and overriding criterion is *historical precedent*: an airline which was assigned a slot in the same previous season ("summer" or "winter") and utilized that slot for at least 80% of the time during that previous season is automatically entitled to the continued use of that "historical slot". *No economic criteria* are used for slot allocation and, in fact, buying and selling of slots is prohibited – at least, under the official rules. The net result is that some of the older, traditional airlines maintain in this way a "lock" on most of the prime slots at the world's most economically desirable airports. Airlines that wish to compete in these markets and may be willing to pay high fees for the right to operate at the airports in question are effectively "frozen out".

Third, and perhaps most important, the RS approach, as currently practiced, distorts the functioning of the marketplace and suppresses potential demand by placing an arbitrary cap on the number of operations at some of the world's busiest airports. This, in turn, creates an illusionary equilibrium, i.e., an artificial balance between demand and capacity. Thus, current RS practices do not provide decision-makers with true information about the economic value that airspace users and the traveling public may attach to additional capacity at the schedule-coordinated airports or at other elements of the airspace.

A great deal of research has been performed over the years on these and related issues, focusing primarily on the second of the problems described above and, to a lesser extent, on the third. Several classical and more recent papers have dealt with the application of "market-based" mechanisms, typically in combination with administrative measures, to improve the current capacity allocation process of IATA – see Vickrey (1969), Carlin and Park (1970), Morrison (1987), Daniel (1995), DotEcon Ltd. (2001), and Ball et al. (2005), for a sample. Such papers examine the use of congestion pricing and of slot auctioning at major airports. Interestingly, some ongoing research on *real-time* capacity allocation and slot-exchange mechanisms in connection with ATFM – see, e.g., Vossen and Ball (2006b) and Ball et al. (2005) – has also focused on market-based approaches (see also Section 4.4).

Other recent work has investigated less generic issues related to computational and implementation issues associated with such approaches. For example, advances in the application of queuing theory have facilitated greatly the estimation of external delay costs (Andreatta and Odoni, 2003). Fan and Odoni (2002) and Hansen (2002) report applications of queuing methodologies to the detailed estimation of external delay costs at New York/LaGuardia and Los Angeles International, respectively. These independent studies come up with strikingly similar results: many flights at these airports impose external delay costs which exceed by at least an order of magnitude the landing fees these flights pay. For example, Fan and Odoni (2002) estimate that the external delay cost per movement for much of the day at New York/LaGuardia was about $6000 in August 2001, whereas the average fee per movement amounted to about $300. The implication is that access to many busy airports may be greatly under-priced, thus attracting excessive demand, which exacerbates congestion. This is especially true in the United States where landing fees are comparatively very low. A third area of recent research addresses some difficult problems that set the application of market-based mechanisms at airports apart from analogous applications in other contexts. For example, Bruckner (2002) and Fan (2003) point out that each airline operates a (possibly large) number of flights at airports, in contrast to highway traffic where each user operates a single vehicle. When any single airline operates a large fraction of the total movements at an airport, it also automatically absorbs ("internalizes") a similarly large fraction of the external delay costs that its flights generate. That airline should therefore be charged only that portion of the external costs that it does not internalize. But a pricing system under which different airlines would pay different landing fees for the same type of aircraft would be both controversial and technically difficult to implement. This and other complications, along with the presence of many social policy objectives (service to small communities, access to airports by regional carriers and by general aviation, etc.) suggest that any future RS schemes that incorporate market-based features, will be governed by a complex set of rules that may include exemptions, subsidies for certain carriers, slots reserved for certain types of flights, etc.

4 Air traffic flow management

4.1 Background and ATFM controls

4.1.1 Basic premises

Air traffic management (ATM) is now viewed as consisting of a *tactical* component and a *strategic* component. The tactical component, Air Traffic Control (ATC), is concerned with controlling individual aircraft on a time horizon ranging from seconds to 30 minutes for the purpose of ensuring safe separation from other aircraft and from terrain. The strategic component, Air Traffic Flow Management (ATFM) works at a more aggregate level and on a time horizon of up to about 12 hours in the United States and 48 hours in Europe. Its objective is to ensure the overall unimpeded flow of aircraft through the airspace, so as to avoid congestion and delays and, when delays are unavoidable, to reduce as much as possible their impact on airspace users.

The primary airports and primary airspace are dominated by the operations of scheduled air carriers. Air carrier schedules generate airspace demand and also serve as the basis for measuring the performance of ATFM systems. The fundamental premise of ATFM is that, roughly speaking, if all operations occurred at their scheduled times and if all airspace elements were in their "normal" operating states, then there would be little need for any flow management. Under such ideal conditions – and with the possible exception of some brief periods at a few of the busiest airports – demand on all airspace elements would be less than capacity and operations would generally proceed as if there were no constraints. But ATFM recognizes that the complexity of the airspace system, its susceptibility to weather conditions and the large number of possible ways in which equipment and/or operations can fail to operate as planned, all imply that the probability that the entire system will operate exactly according to schedule on any given day is essentially zero. ATFM procedures are therefore in greatest demand when there are significant imbalances between capacity and demand on airspace elements, usually caused by capacity reductions due to weather or equipment failures. Demand surges can also cause capacity-demand imbalances. Such surges may occur, for example, when problems early in the day cause the postponement of scheduled operations into a time interval that already contained significant numbers of operations. Additionally, the number of unscheduled flights is growing and is having a greater impact on overall system performance. Finally, as the number of scheduled operations has increased, there have been instances where scheduled demand actually exceeded the capacity of network elements for extended time periods each day. Probably the most notable example along these lines occurred at LaGuardia airport between May 2000 and February 2001, i.e., during the period between enactment of the so-called Air-21 legislation – which opened access to LaGuardia for certain types of flights – and the imposition of slot lotteries aimed at relieving the resulting congestion. In such cases, even the routine operation of the air transportation system requires the use of ATFM procedures.

4.1.2 Controls

ATFM performance is measured primarily with reference to deviations from schedules. That is, ATFM systems generally seek to minimize the amount of time by which actual operations (most importantly, the arrivals of aircraft at their destination airports) deviate from scheduled operations. Thus, the key performance indicators usually involve measures of delay. The next fundamental question to consider in describing and analyzing ATFM systems is what controls can be used to impact system performance. Before directly addressing this question we discuss two basic characteristics of controls: who makes and implements the control decisions and what are the timing constraints associated with the control. The two critical entities involved in ATFM decisions are the air traffic service providers and the airspace users. Airspace users could range from the owner/operator of a general aviation aircraft to a large air carrier. Within a large air carrier there can be multiple decision-makers: most central to this discussion are the airline operational control centers (AOCC) and the pilots. On the air traffic service provider side, the units involved are the regional air traffic control centers (Air Route Traffic Control Centers – ARTCC in the US), as well as the central traffic flow management units (Air Traffic Control System Command Center – ATCSCC in the US). Properly distributing decision-making among all these entities can be critical to the success of ATFM systems and is at the heart of the recent emergence of the Collaborative decision making (CDM) paradigm, which is described in Section 4.3. The time constraints associated with control actions can greatly impact the manner in which they can be applied and the manner in which multiple actions can be coordinated. The key issue, in this respect, is the length of time that elapses between a control decision and the implementation of that decision. A rough categorization is that *strategic decisions* are typically made hours in advance of implementation, whereas *tactical decisions* involve shorter time scales. Generally, the appropriateness and effectiveness of any particular control action, strategic or tactical, depend on system dynamics and on the level of uncertainty associated with future states of the airspace system.

We now describe the most important types of ATFM control actions.

Ground holding (including ground stops). Ground holding involves delaying the departure of a flight in order to avoid overloading a capacitated system element. Ground holding is most often implemented in the US through a Ground delay program (GDP), which is put into effect when the demand for arrivals into an airport is predicted to exceed significantly the arrival capacity. In Europe, ground holding is commonly used to avoid overloading of en-route sectors. Ground holding is generally considered a strategic decision. A "ground stop" is a more tactical and extreme form of ground holding, whereby all departures of aircraft bound for a particular destination airport are temporarily postponed. A ground stop typically applies only to a specified set of airports, usually ones that are proximate to the affected destination airport. In the past, air traffic service providers were the sole decision makers when it came to de-

cisions regarding ground holding. More recently, however, CDM procedures have led to shared decision making with airspace users.

Airborne speed control and airborne holding. Airborne speed control and airborne holding are tactical controls used to avoid overloading (en-route or terminal) airspace elements by adjusting the time at which flights arrive at those elements. Speed controls can involve simply slowing down or speeding up aircraft or aircraft vectoring implemented through minor directional detours. Airborne holding is usually implemented by having aircraft fly in oval-shaped patterns. The providers of air traffic services usually make these control decisions.

Route choice and route adjustments. Generally speaking airspace users control route choice. As discussed at the beginning of Section 2, airline and general aviation operators choose and file flight plans based on several criteria including winds and other weather conditions, fuel usage, en-route turbulence predictions, safety constraints, etc. Flight plans may be filed several hours in advance of departure but in many cases are not filed until shortly before departure (even within an hour of departure) in order to take advantage of the most recent information on weather conditions and congestion. For purposes of managing congestion, air space managers can reject flight plans leading to new filings. In some cases, standard reroute strategies are specified, e.g., the air space manager designates that all flights originally filed along one airspace path should re-file along a second alternate path. A variety of more tactical route adjustments are possible. For example, alternative "departure routes" might be specified by an airline; then, immediately before departure, one would be chosen based on a negotiation process between the airspace manager and user. Once airborne, major or minor route adjustments might be made. These decisions generally involve some level of collaboration between the regional air traffic service provider and the pilot or AOCC. A very extreme form of route adjustment is the *diversion* of a flight, which involves altering its destination airport.

Flight cancellations. The cancellation of a flight is another drastic, although not particularly unusual, ATFM measure. Responsibility for flight cancellations always rests with the airline or the general aviation operator.

Arrival sequencing. The sequencing of flights can be very important as the maximum arrival rates into airports depend on the sequence and mix of aircraft types (Section 2.1.3). Due to the uncertainty of en-route operations this should be considered a tactical decision. Primary responsibility in this respect rests with the air traffic service provider, although CDM concepts are now being applied to this setting, leading to some air carrier participation in sequencing decisions.

Airport arrival or departure rate restrictions. As discussed in Section 2, there can be a tradeoff between the maximum airport arrival and departure rates. Evaluating this tradeoff and setting these rates should be considered a tactical decision that traditionally is made by regional air traffic service providers (Gilbo, 1993). Proposals to allow air carrier participation in this decision have been made (Hall, 1999).

Waypoint flow restrictions. An important and widely used control within the US airspace system is the so-called *miles-in-trail* (MIT) restriction. Regional air traffic service providers impose such restrictions in order to ensure that the flow of aircraft into a region of airspace is kept at a safe level. When such a restriction is put in place, the adjacent "upstream" regional air traffic service provider has responsibility for maintaining a traffic flow at or below the restricted level. This can be done in a variety of ways, including the use of airborne holding, rerouting of some traffic and issuing similar flow restrictions on flights further upstream. In this way, it is possible for a flow restriction to propagate through much of the airspace system, possibly eventually leading to ground holds at airports of origin.

4.2 Deterministic models

Using as a starting point the capacity constraints described in Section 4.1, one can formulate large-scale optimization models that map airspace demand onto the various elements of the airspace in such a way that all capacity constraints are respected. Such models trace flights through both space and time and seek to minimize some global demand function. In Section 4.2.1 we describe a large-scale comprehensive modeling approach and then, in Section 4.2.2, specialize this model to the ground holding problem.

4.2.1 Global models

There are two broad classes of global air traffic flow models. The first assumes that the trajectory (route) of each flight is fixed and optimizes the timing of the flight's operations. The second allows the route of each flight to vary, as well. Clearly, the second type of model has a much larger decision space.

We start with models of the first type. The modeling approach chooses a time horizon of interest and decomposes it into a discrete set of time intervals. A geographic scope is also selected. This determines the set of capacitated elements under consideration, including airports, sectors, and waypoints. The combination of the model's temporal and geographic scope determines the set of flights to be considered. The basic decision variable specifies the airspace element occupied by a flight at each time interval, i.e.,

$$x_{fte} = \begin{cases} 1 & \text{if flight } f \text{ occupies airspace element } e \\ & \quad \text{during time interval } t; \\ 0 & \text{otherwise.} \end{cases}$$

For airports, the capacitated airspace element would be either the airport's arrival or departure component. For example, LaGuardia Airport (LGA) would be represented by an arrival element $LGA1$ and a departure element $LGA2$, with $x_{f1,t,LGA1}$ indicating the time interval t of a flight $f1$'s arrival to LGA and $x_{f2,t,LGA2}$ indicating the time interval t for flight $f2$'s departure from LGA.

The capacity constraint associated with an element e and time interval t is of the form:

$$\sum_f x_{fte} \leqslant \text{cap}(t, e) \quad \text{for all } t \text{ and } e,$$

where $\text{cap}(t, e)$ is the capacity of element e during time interval t. For airport arrival and departure capacities and for waypoints, $\text{cap}(t, e)$ is equal to the maximum number of flights that could flow through that element during time interval t. For a sector, it is equal to the maximum number of flights that can occupy the sector simultaneously.

The remaining constraints define temporal restrictions on the manner in which each flight can progress through the airspace. For example, they might specify that, once a flight enters a sector, it must remain in the sector for 3, 4, or 5 time intervals. In this case, 3 time intervals would correspond to traversing the sector on a direct path at maximum speed and 5 time intervals might correspond to a longer traversal time based on application of some type of speed control. We note that since the flight's route is an input, the progression from departure airport through a specific sequence of sectors to a destination airport is a fixed model input, as well.

Bertsimas and Stock Paterson (1998) have shown that models of this type can be solved very efficiently. Of particular note is their use of an alternative set of decision variables. Specifically, the x_{fte} variables are replaced with a set, w_{fte}, defined by

$$w_{fte} = \begin{cases} 1 & \text{if flight } f \text{ arrives at airspace element } e \text{ by time interval } t; \\ 0 & \text{otherwise.} \end{cases}$$

While the w variables can be obtained from the x variables through a simple linear transformation ($w_{fte} = \sum_{s=1}^{t} x_{fse}$), the w variables are much easier to work with because they produce very simple and natural temporal progression constraints. Further, the associated linear programming relaxations are very "strong" in the sense that they lead to the fast solution of the associated integer programs. Bertsimas and Stock Patterson show that a variety of additional features can be included in the model, including the propagation of delays that occurs when a delay in the arrival of a flight arrival causes the delay of an outbound flight that uses the same aircraft. The model can also capture the dependence between an airport's arrival and departure capacities and choose the appropriate combination of the two for each time interval.

A second type of model also allows for flight routes to vary. Bertsimas and Stock Paterson (2000) extend the model described above to this case. Since the decision space becomes much larger, aggregate variables and approximate methods must be employed in order to solve problems of realistic size.

4.2.2 Deterministic ground holding models

A simple, yet very important special case of the model just described is the deterministic ground holding problem. For this problem, en-route constraints and airport departure constraints are ignored leaving only a constraint on arrival capacity. These assumptions allow a large multi-airport problem to be decomposed into separate problems, one for each arrival airport. This model is very important to the US ATFM environment since it underlies the Ground delay programs (GDP) as currently operated. Although the control variable for a GDP is ground delay at the origination airport of each flight, the problem can be modeled as one of assigning flights to arrival time intervals at the destination airport. For each flight a (constant) en-route time is then subtracted from the arrival time to obtain a departure time, which in turn implies an amount of ground delay at the origination airport.

Define the following inputs: for a set of time intervals t ($t = 1, 2, \ldots, T$) and set of flights f ($f = 1, 2, \ldots, F$), let

$$
\begin{aligned}
b_t &= \text{the constrained airport's arrival capacity at time interval } t,\\
e(f) &= \text{the earliest time interval at which flight } f\\
&\qquad \text{can arrive at the constrained airport,}\\
c_{ft} &= \text{the cost of assigning flight } f \text{ to arrive at time interval } t,
\end{aligned}
$$

and the variable

$$
x_{ft} = \begin{cases} 1 & \text{if flight } f \text{ is assigned to time interval } t;\\ 0 & \text{otherwise.} \end{cases}
$$

Then, the deterministic ground holding problem can be formulated as

$$
\begin{aligned}
\min \sum_f \sum_t c_{ft} x_{ft}\\
\text{subject to} \quad \sum_f x_{ft} &\leqslant b_t \quad \text{for all } t,\\
\sum_{t \geqslant e(f)} x_{ft} &= 1 \quad \text{for all } f,\\
x_{ft} &\geqslant 0 \text{ and integer} \quad \text{for all } f \text{ and } t.
\end{aligned}
$$

As can be seen, this is a simple transportation model that can be solved very efficiently. This model was first described in Terrab and Odoni (1993). In Ball et al. (1993), various issues related to its practical use were investigated. In particular, it was suggested that the definition $c_{ft} = r_f(t - e(f))^{1+\varepsilon}$ is attractive since flight delay costs tend to grow with time at a greater than linear rate. In addition, solutions produced using this objective function are attractive from the standpoint of equity or fairness – see Vossen and Ball (2006a) for a detailed discussion of this issue. Hoffman and Ball (2003) investigate generalizations of this model that preserve the proximity of banks of flights associated with airline

hubbing operations. Vranas et al. (1994) describe a multi-airport version of the problem where both airport arrival and departure rates are constrained and flights are "connected" to each other, in the sense that the late arrival of a flight f at an airport may also imply the late departure of one or more subsequent flights from that airport. This can happen if the departing flight must use the same aircraft as flight f or, less obviously, if flight f carries many transfer passengers who must board subsequent departing flights.

4.3 Uncertainty and stochastic models

Uncertainty on multiple levels has led to the failure of many attempts at practical implementation of various air traffic flow management models. This is particularly true for models that attempt to optimize over a broad geographic area and/or extended periods of time. To be effective, models must include stochastic components explicitly or they must address problems restricted to limited geographic and time domains for which available information is less subject to uncertainty. We use the term *demand uncertainty* to describe the possibility that, due to random or unforeseen events, flights may deviate from their planned departure or arrival times or from their planned trajectories. Similarly, *capacity uncertainty* refers to the possibility that random or unforeseen events will cause changes to the maximum achievable flow rates into and out of airports or through airspace waypoints or to the maximum number of flights that can occupy simultaneously a portion of the airspace. Examples of factors contributing to demand uncertainty include problems in loading passengers onto an aircraft, mechanical problems, queues on the departure airport's surface or in the air, and en-route weather problems. Examples of factors contributing to capacity uncertainty include weather conditions at an airport or in the en-route airspace, failures or degradation of air traffic control equipment, unavailability of air traffic control personnel and random changes in flight sequences, and aircraft mix that require alterations of anticipated flight departure or arrival spacing.

The largest body of work in this area has focused on ground holding models that explicitly take into account uncertainty in airport arrival capacity. These models are covered in Section 4.3.1. There is a less extensive body on work on capacity uncertainty for GDPs, which is covered in Section 4.3.2.

The performance of a GDP planning and control system can be evaluated based on three measures: the total assigned ground delay, the total airborne delay and the utilization of the available arrival capacity. Figure 5 illustrates a generic model (Odoni, 1987) that is convenient for evaluating the tradeoffs among these three performance criteria and for understanding how they are impacted by demand and capacity uncertainty. The airport's arrival component (terminal airspace and runways) is viewed as a server subject to demand uncertainty. The assigned ground delay (the control mechanism) and the random events that perturb the planned arrival stream of flights determine the rate of arrivals at the server. Thus, demand uncertainty impacts the rate of

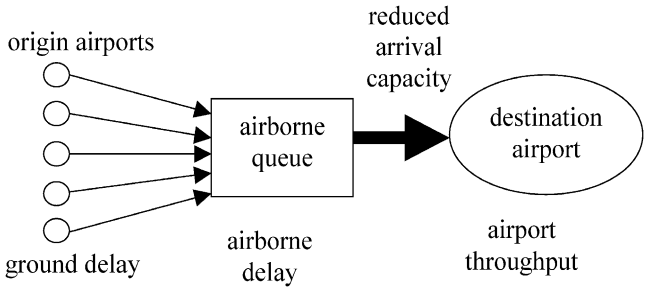


Fig. 5. Schematic representation of a single-airport GDP.

arrivals at the server and capacity uncertainty impacts the service rate. The planning and control of a GDP must balance the possibility of too low an arrival rate, which leads to underutilization of the server, with too high an arrival rate, which leads to a large airborne queue and excessive airborne delay. Generally speaking because of demand and capacity uncertainty, the "best" policy calls for some degree of airborne delay in order to ensure an acceptably intensive utilization of the available arrival capacity. The starting point for defining airport arrival capacity is specifying the number of flights that can land within a time interval (this quantity is referred to as the airport acceptance rate – AAR). A key observation in light of this discussion is that the presence of demand and capacity uncertainty makes it necessary to distinguish between the planned AAR (PAAR) and the actual AAR. In the following two sections we discuss two types of models: those that assign ground delay to individual flights and those only concerned with determining a PAAR vector. The second objective (determining a PAAR vector) is appropriate in the context of Collaborative Decision Making (CDM), where the assignment of ground delays to individual flights results from a complex set of distributed decision making procedures. Thus, the second set of models is the one more directly geared to the CDM environment. Interestingly, some models in the first set can also generate the same PAAR vectors through a simple "post-processing" step, as has been shown by Kotnyek and Richetta (2004).

4.3.1 Ground holding models with stochastic airport capacity

As discussed in the previous section, optimization models for the ground holding problem subdivide time into an arbitrary number of discrete intervals. Typical time intervals might be 10 or 15 minutes or even as much as 1 hour for the most aggregate models. One would then characterize the AAR over a period of time, e.g., four hours, as a vector that provides the AAR in each of the constituent time intervals. In a typical GDP caused by a weather disturbance that moves through a region, the AAR would start at its normal level, e.g., 60 arrivals per hour, decrease to one or more degraded levels, e.g., 30 arrivals per hours, for several time intervals, e.g., 4 hours, and then return to its original level. If it were known that such a scenario would occur with certainty,

then a deterministic ground holding model (Section 4.2.2) could obviously be used with this scenario providing the capacity constraints input. By contrast, stochastic models treat the AAR as a random variable and use as input several such scenarios together with associated probabilities. For example, an "optimistic" scenario indicating no capacity degradation would consist of a vector of hourly AARs of 60 throughout the period of interest, whereas, a more pessimistic scenario might assume that the AAR will be 30 during every hour in the period. The input can thus be characterized as

$$\begin{aligned} D_{tq} &= \text{AAR for time interval } t \text{ under scenario } q, \\ &\qquad \text{for } t = 1, \ldots, T \text{ and } q = 1, \ldots, Q. \\ p_q &= \text{ probability of the occurrence of scenario } q, \text{ for } q = 1, \ldots, Q. \end{aligned}$$

The x_{ft} variables are then defined as in the deterministic ground holding model, but the capacity constraints are replaced with a new set of scenario-based constraints and associated variables. The new variable set is defined by

$$\begin{aligned} y_{tq} &= \text{the number of flights held in the air from time period } t \text{ to } t+1, \\ &\qquad \text{under scenario } q, \text{ for } q = 1, \ldots, Q. \end{aligned}$$

The new set of capacity constraints then

$$\sum_{f=1,F} x_{ft} + y_{t-1q} - y_{tq} \leqslant D_{tq} \quad \text{for } t = 1, \ldots, T \text{ and } q = 1, \ldots, Q.$$

Thus, under these constraints, there is a separate capacity for each scenario. However, the y variables allow for flow between time intervals, so the number of flights assigned to land in an interval under a particular scenario can exceed the AAR by allowing excess flights to "flow" to a future time interval. Note that this set of capacity constraints defines a small network flow problem for each q, with flights "flowing" from earlier time intervals to later ones. To be feasible, for each given q, the total arrival capacity for the entire period of interest, $\sum_t D_{tq}$, must be at least as large as the total number of flights (F).

The objective function for the model minimizes the sum of the cost of ground delay plus the expected cost of airborne delay. If we define c^{a} as the cost of holding one flight in the air for one time period then we can define the objective function as

$$\min \sum_{f=1,F} \sum_{t=1,T} c_{ft} x_{ft} + \sum_{q=1,Q} p_q \sum_{t=1,T} c^{\mathrm{a}} y_{t-1q}.$$

This class of model was first described in Richetta and Odoni (1993) and a dynamic version was given in Richetta and Odoni (1994). Ball et al. (2003) defines a simpler version of this model that computes a PAAR vector (i.e., the total number of flights assigned to arrive during each time interval in the period) rather than an assignment of ground delay to individual flights or groups

of flights. This model was created in order to be compatible with CDM procedures, which will be discussed in Section 4.4. The constraint matrix of the underlying integer program is the transpose of a network matrix, allowing the integer program to be solved using linear programming or network flow techniques. Inniss and Ball (2002) and Wilson (2004) describe recent work on estimating arrival capacity distributions. Willemain (2002) develops simple GDP strategies in the presence of capacity uncertainty and also takes into account the possibility of flight cancellations. Inniss and Ball (2002) also presents a dynamic approach to this problem.

4.3.2 Modeling demand uncertainty

Recent experience with new decision support tools for planning and controlling GDPs has strongly suggested that demand uncertainty is the reason that these tools have not performed quite as strongly as expected. In this context, demand uncertainty refers to differences between the planned and actual characteristics of the stream of flights arriving at the destination airport. These differences can be attributed to three categories of causes:

(1) cancellations – flights in the planned arrival stream that are absent from the actual arrival stream during the planning time period;

(2) "pop-ups" – flights that arrive at the destination airport during the planning time period, but were not in the planned arrival stream;

(3) "drift" – a discrepancy between the actual arrival time of a flight and the flight's planned arrival time.

Ball et al. (2001c) present a demand uncertainty model, where each flight can be canceled with a known probability, pop-ups arrive in each time interval according to a binomial distribution and flights drift within predefined intervals according to a uniform distribution. Considering only cancellations and pop-ups, the authors provide an optimization model that determines a PAAR that minimizes airborne delay subject to a constraint on the minimum allowable airport capacity utilization. This integer programming model contains embedded binomial distributions. The authors also present simulation results covering all three forms of demand uncertainty. The results not only provide new PAAR setting policies, but also give an approach to estimating the benefits of reducing demand uncertainty. For additional details see Bhogadi (2002) and for a related model Willemain (2002).

4.4 Collaborative decision making

The Collaborative decision making (CDM) (effort Ball et al., 2001a, 2001b; Chang et al., 2001; Wambsganss, 1996) grew out of a desire on the part of both the airlines and the FAA for improvements in the manner in which GDPs were planned and controlled. The FAA and, more specifically, the ATCSCC had realized the need for more up-to-date information on the status of flights currently delayed due to mechanical or other problems or even canceled unbeknownst to the ATCSCC. Equally important, the success of GDPs also depends

vitally on timely information regarding airline intentions vis-à-vis flight cancellations and delays over the next few hours. At the same time, the airlines did not feel the allocation procedures used by the ATCSCC were always fair and efficient. In addition, each airline wished to gain more control over how delays were allocated among its own flights. Thus, both the airlines and the FAA had specific (although different) objectives that motivated their participation in CDM.

The CDM "philosophy", broadly speaking, represents an application of the principles of information sharing and distributed decision making to ATFM. Specific goals are:

- generating a better "knowledge base" by merging information provided by the airspace users with the data that are routinely collected by monitoring directly the airspace;
- creating common situational awareness by distributing the same information to both traffic managers and airspace users;
- creating tools and procedures that allow airspace users to respond directly to capacity/demand imbalances and to collaborate with traffic flow managers in the formulation of flow management actions.

4.4.1 CDM resource allocation procedures for ground delay programs

CDM brings new modeling requirements to ATFM resource allocation problems including: considering allocation at a more aggregate level, e.g., by allocating resources to airlines rather than to individual flights; integrating equity criteria into model objectives and/or constraints; ensuring that resource allocation mechanisms provide incentives for information sharing; and developing inter-airline resource exchange mechanisms. These concepts evolved in the process of developing new allocation methods for GDPs. We describe the GDP procedures here, as well as related recent research and efforts to extend these ideas to en-route ATFM problems.

As discussed in Section 4.2.2, the GDP planning problem can be viewed as one of assigning each flight in the GDP to an arrival time interval (or time slot). Figure 6 illustrates the GDP resource allocation process under CDM. The FAA, using *Ration-by-schedule* (RBS), provides an initial assignment of slots to flights. Each airline, using the *cancellations and substitution* process, may then cancel flights and modify slot-to-flight assignments for its own flights (*intra*-airline exchange). Thus, although RBS, in concept, allocates slots to flights, the cancellation and substitution process effectively converts the slot-to-flight

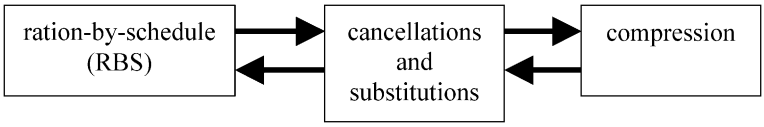


Fig. 6. CDM resource allocation procedures.

assignment into a slot-to-airline assignment. The final step, *compression*, which is carried out by the FAA, maximizes slot utilization by performing an *inter*-airline slot exchange in order to ensure that no slot goes unused.

The assignment of time slots by RBS can be viewed as a simple priority rule. Using the *scheduled* arrival order as a priority order, each flight is assigned the next available arrival slot. If this rule were applied to all flights and there were no cancellations or substitutions, then the flights would arrive in their original sequence, but generally later in time. Two groups of flights are exempted from this basic allocation scheme:

(1) flights that are currently airborne (clearly these cannot be assigned any ground delay) and

(2) a set of flights selected according to the distance of their departure airports from the GDP (arrival) airport (Ball and Lulli, 2004).

The motivation for (2) is to include in the allocation scheme flights originating from airports that are close to the GDP airport and to exempt flights from airports further away from the GDP airport. The reasoning is that flights a greater distance away must be assigned ground delays well in advance of their actual arrival at the GDP airport, e.g., 4 or 5 hours in advance. At such a long time before arrival, there tends to be a greater level of uncertainty regarding weather and, as a consequence, airport arrival capacity. If these distant flights are assigned ground delay, there is a significant likelihood that this ground delay might prove unnecessary. Thus, distance-based exemptions constitute a mechanism for avoiding unrecoverable delay, as well as for increasing expected airport throughput.

After the round of substitutions and cancellations, the utilization of slots can usually be improved. The reason for this is that an airline's flight cancellations and delays may create "holes" in the current schedule; that is, there will be arrival slots that have no flights assigned to them. The purpose of the compression step is to move flights up in the schedule to fill these slots. The basic premise behind the algorithm currently used to perform compression is that airlines should be "paid back" for the slots they release, so as to encourage airlines to report cancellations.

To illustrate the compression algorithm, consider the example shown in Figure 7. The left side of the figure represents the flight-slot assignment prior to the execution of the compression algorithm. Associated with each flight is an earliest time of arrival, and each slot has an associated slot time. Note that there is one canceled flight. The right side of the figure shows the flight schedule after execution of the compression algorithm: as a first step, the algorithm attempts to fill AA's open slot at 1602–1603. Since, there is no flight from AA that can use it, the slot is allocated to UA, and the process is repeated with the next open slot, which, using the same logic, is assigned to US. The process is repeated for the next open slot, which is now assigned to AA. AA thus receives the earliest slot that it can use. The net result of compression in this case can be viewed as an exchange among airlines of the slots distributed through the initial RBS allocation.

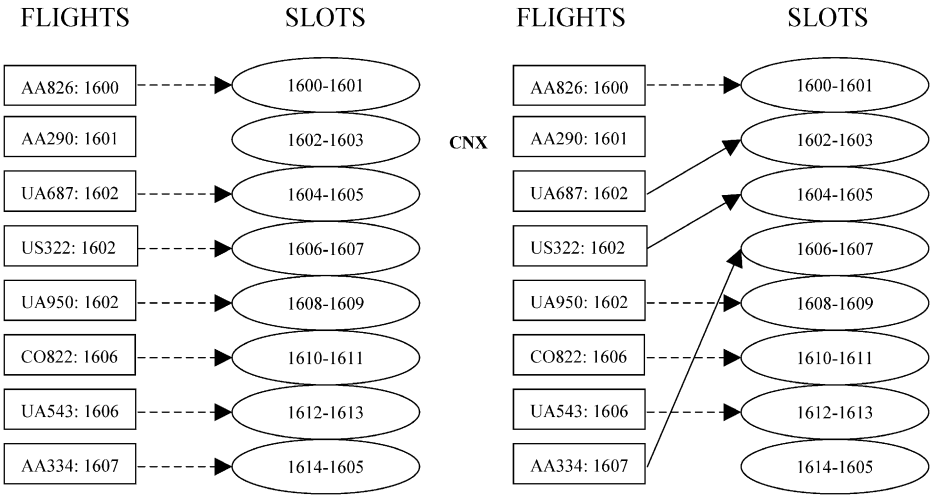


Fig. 7. Execution of the CDM compression algorithm.

4.4.2 CDM concepts, philosophy, and research directions

We now investigate the special properties of the CDM resource allocation procedures and describe how these are being extended and enhanced.

RBS, compression and information sharing. One of the principal initial motivations for CDM was that the airlines should provide updated flight status information. It was quickly discovered that the existing resource allocation procedures, which prioritized flights based on the current estimated time of arrival, actually discouraged the provision of up-to-date information. This was because, by updating flight status, the airlines would change the current estimate of flight arrival (almost always to a later time), which in turn changed their priority for resource allocation. On the other hand, the RBS priority, which is based on the (fixed) schedule, does not vary with changes in flight status. Further, as in the example of Figure 7, compression most often provides an airline with another usable slot whenever an announced cancellation generates a slot that the airline cannot fill via the substitution process.

Properties of RBS. We summarize some basic properties of RBS derived in Vossen and Ball (2006a). RBS assigns to each flight, f, a controlled time of arrival, $CTA(f)$. This is equivalent to assigning a delay, $d(f)$, to flight, f, given by $d(f) = CTA(f) - OAG(f)$, where $OAG(f)$ is the scheduled arrival time for f. All time values are rounded to the nearest minute under RBS, hence, each delay value $d(f)$ is integer. If we let D equal the maximum delay assigned to any flight and $a_i = |\{f\colon d(f) = i\}|$ for $i = 0, 1, 2, \ldots, D$, then the important properties of *unconstrained RBS* (RBS with no flight exemptions) can be defined by the following properties.

Property 1. RBS minimizes total delay $= \sum_f d(f)$.

Property 2. RBS lexicographically minimizes $(a_D, \ldots, a_1, a_0)$. That is, a_D is minimized; subject to a_D being fixed at its lexicographically minimum value, a_{D-1} is minimized; subject to (a_D, a_{D-1}) being fixed at its lexicographic minimum value, a_{D-2} is minimized; and so on.

Property 3. For any flight f, the only way to decrease a delay value, $d(f)$, set by RBS is to increase the delay value of another flight g to a value greater than $d(f)$.

CDM and equity. Property 3, which follows directly from Property 2, expresses a very fundamental notion of equity that has been applied in a number of contexts (Young, 1994). It is remarkable that procedures, such as RBS, which were developed in very practical war-gaming and consensus-building exercises have such elegant and desirable properties. On the other hand, this may not be surprising in that these properties represented the basis for reaching consensus, in the first place. The properties show that unconstrained RBS produces a fair allocation. However, one should consider whether the exemption policies described earlier, in fact, introduce bias. Vossen et al. (2003) show that exemptions do introduce a bias and also describe procedures for mitigating these biases. The approach taken initially computes the unconstrained RBS solution and defines it as the "ideal" allocation. Optimization procedures are then described that minimize the deviation of the actual allocation from the ideal. These procedures build upon approaches developed in connection with just-in-time (JIT) manufacturing for minimizing the deviation of an actual production schedule from an ideal schedule – see, for example, Balinski and Shihidi (1998). The resulting approaches maintain the exemption policies, but take into account the "advantages" provided to an airline by its exempted flights when allocating delays to its other flights.

Compression as trading. Although the initial interpretation of compression is as a slot reallocation procedure that maximizes slot utilization, there is also a natural interpretation of compression as an inter-airline trading or bartering process (Vossen and Ball, 2006a). For example, in Figure 7, American Airlines "traded" the 1602–1603 slot, which it could not use for the 1606–1607 slot, which it could use, and United Airlines reduced its delay by trading the 1604–1605 slot for the earlier 1602–1603 slot. Vossen and Ball show that a bartering process can be structured so as to produce a result essentially equivalent to compression. This view of compression suggests many possible extensions. For example, Vossen and Ball (2006b) define a more complex 2-for-2 bartering mechanism and show that using this mechanism offers a substantial potential for improved economic performance. Probably the most intriguing enhancement is allowing "side payments" with any exchange as well as the buying and

selling of slots. Ball et al. (2005) provide a discussion of this and other aviation-related market mechanisms.

4.5 Air carrier response options

As the implementation of the CDM concept expands, it will become increasingly difficult to designate certain controls as exclusive to the traffic managers or exclusive to the users (airlines). In this section we describe the types of actions and types of problems that air carriers can take in response to ATFM initiatives based on current practices. We start by noting that the decision on whether or not a flight takes place on a given day rests with the airlines, with the exception of extreme circumstances. Further, the airlines generally control the route choice for a flight subject to constraints on access to portions of airspace exercised by ATFM. The airlines can, of course, at their own discretion choose to delay the departure of flights. On the other hand, ATFM exercises strong control over the timing of operations. As described in the previous section, for GDPs in the US, ATFM also exercises control over flight timing by allocating arrival time slots to airlines, which in turn allocate the slots to individual flights.

For the case of GDPs, we can model an airline's response problem as one of assigning flights to the set of time slots that airline "owns". A first-level approach to this problem would select an assignment that minimizes the "cost" of the associated flight delays. Such models should also consider the possibility and cost of canceling flights. Second-level models would also consider the downstream effects of the delayed flights. For example, the immediate downstream effect involves the delays on flights and passengers outbound from the GDP airport (Niznik, 2001). Broader airline network models would consider the airport arrival slots as a resource to be allocated as part of a larger optimization model.

The problem of choosing a route is traditionally addressed on a flight-by-flight basis. There are a variety of safety regulations that constrain this problem. Typically, weather conditions, including winds, turbulence, convective activity, etc., play a strong role in solving this problem. Airspace congestion considerations, as well as ATFM control actions are also playing an increasingly important role. There can be a high degree of uncertainty in many of these factors leading to the need for stochastic, dynamic problem solving approaches – see, for example, Nilim et al. (2001). Furthermore, as CDM concepts become pervasive, the air carriers may be allocated limited airspace resources that in turn need to be allocated to individual flights. Under such a scenario, flight-by-flight route planning may no longer be viable. Berge et al. (2003) describe a comprehensive optimization model that includes decision variables for the three controls described earlier, i.e., flight cancellations, aircraft route choice, and ground delay. It assumes an environment in which airspace and airport resources have been allocated to each airline. This model was developed for integration into a comprehensive Boeing airspace

simulation. It can, however, also be viewed as a prototype for future operational airline decision models.

Sections 6 and 7 describe models for planning adjustments to airline operations based on day-to-day changes in airspace conditions and airline resource status. Such models typically are only invoked in the case of reasonably significant disturbances to "normal" conditions and the underlying environment is called *irregular operations*. While these models take into account some of the concepts described in this section, a full integration of the emerging ATFM CDM philosophy with airline planning models represents a research challenge.

5 Simulation models

Simulation models are useful support tools in understanding and visualizing the impact of certain types of disruptive events on airport, airspace, and airline operations and on air traffic flows, as well as in testing the effectiveness of potential responsive actions. This section reviews briefly some of these tools.

For analyzing impacts at the local airport or airspace level or for testing the effectiveness of tactical ATFM actions, simulation models need to be highly detailed. When it comes to questions at a regional or strategic level, models of a more aggregate nature are often more appropriate. Detailed (or "microscopic") simulation models of airport and airspace operations first became viable in the early 1980s and have been vested with increasingly sophisticated features since then. The models that currently dominate this field internationally are the Airport and Airspace Simulation Model (better known as SIMMOD), the Total Airport and Airspace Modeler (TAAM), and the Reorganized ATC Mathematical Simulator (RAMS). The first two are widely used in studies dealing with the detailed planning and design of airports and/or of volumes of airspace. SIMMOD is publicly available through the FAA, but more advanced proprietary versions can be obtained through private vendors (see, e.g., ATAC, 2003). TAAM is a proprietary model (Preston, 2002). Finally, RAMS (Eurocontrol, undated) is also a proprietary model limited to detailed simulation of airspace operations and procedures and is used either as a training tool for air traffic controllers or for studies of controller workload in airspace sectors. To our knowledge, none of these, or any other, detailed simulation models is currently utilized on a routine basis as a "real time" support tool for the types of dynamic tactical or strategic ATFM actions described in Section 4. However, the models have reached a state of development where such use is technically feasible. For example, in a case where a runway at an airport is temporarily out of use (e.g., due to weather conditions or to a disabled aircraft), one could simulate and compare the effectiveness of alternative allocation schemes for the assignment of arrivals and departures to other runways or of various runway-use sequences. Odoni et al. (1997) provides detailed – and somewhat dated, by now – descriptions of several microscopic simulation models, including the three mentioned above.

Another class of models has been developed to support more aggregate analysis typically involving a broader scope than the problems addressed by SIMMOD or TAAM. For example, NASA has sponsored the development of FACET (Future ATM Concepts Evaluation Tool – Bilimoria et al., 2000). Typical uses might involve analyzing the impact over the entire national airspace of new traffic flow management initiatives, air-ground distributed control architectures, and decision support tools for controllers. FACET models system-wide en-route airspace operations over the contiguous United States. It strikes a balance between flexibility and fidelity enabling the simultaneous representation of over 5000 active flights on a desktop computer.

The CDM activities have also required the extensive use of human-in-the-loop (HITL) experiments in order to test new distributed decision-making ideas. These experiments were initially supported by one-of-a-kind computer-communications architectures. More recently the FAA has funded the development by Metron Aviation of the Jupiter Simulation Environment (JSE). JSE can emulate the message stream generated by the FAA's Enhanced Traffic Management System (ETMS). ETMS provides real-time information on the status of all flights operating within the US airline operational control centers and FAA facilities can connect their traffic monitoring systems to the JSE and participate in HITL experiments involving new decision support tools or operational concepts. For example, the JSE allowed for the rapid testing and evaluation of the slot credit substitution (SCS) concepts and tools prior to their release.

Simulation models are equally important on the airline operations side. To evaluate the recovery procedures and plans for fleets and crews under operational conditions, it is necessary to have them simulated by a model of airline operations. MEANS (Clarke et al., 2004) and SIMAIR (Rosenberger et al., 2002) are such simulations. The MIT Extensible Air Network Simulation (MEANS) can be used to predict the effects of air traffic control, traffic flow management, airline operations control, and airline scheduling actions on air transportation system performance, measured in terms of airport congestion and throughput, and aircraft, crew, and passenger delay, and disruption. MEANS has a modular architecture and interface, with each module corresponding to a set of operational decisions made by air transportation coordinators and controllers. For example, flight cancellation and re-routing decisions are made in the Airline Module; the amount of traffic allowed between airports is determined by the Traffic Flow Management Module; and the sequencing and spacing of aircraft is performed by the Tower/TRACON Module. This modular structure provides flexibility for implementing additional and/or more complex modules without requiring changes to the core interfaces.

Another stochastic model of airline operations, SIMAIR, uses an Event Generator module to generate events such as arrivals, departures and repaired planes. The generator samples random ground time delays, block time delays, and unscheduled maintenance. SIMAIR contains two modules for decision-making. The Controller module maintains the state of the system. It emulates

an Airline Operational Control Center in the sense that it recognizes disruptions and implements recovery policies. If a disruption prevents a leg from being flown as planned, the Controller requests a proposed reaction from the Recovery Module, which it can accept or request another, time permitting.

SIMAIR has been used to evaluate the recovery procedures and robust plans presented in several of the papers discussed in Sections 6 and 7. A SIMAIR Users Group is in place, consisting of several research groups and airlines that have used or contributed to the development of SIMAIR.

6 Airline schedule recovery

When disruptions occur, airlines adjust their flight operations by delaying flight departures, canceling flight legs, rerouting aircraft, reassigning crews or calling in new crews, and re-accommodating passengers. The goal is to find feasible, cost-minimizing plans that allow the airline to recover from the disruptions and their associated delays. To address this challenge, airlines have established Airline Operational Control Centers (AOCC) to control safety of operations, exchange information with regulatory agencies, and manage aircraft, crew, and passenger operations. The AOCC is comprised of (Bratu, 2003; Clarke et al., 2000; Filar et al., 2000):

- *Airline operations controllers* who, at the helm of the AOCC, are responsible for aircraft re-routing, and flight cancellation and delay decisions for various types of aircraft.
- *Crew planners* who find efficient recovery solutions for crews and coordinate with airline operations controllers to ensure that considered operations decisions are feasible with respect to crews.
- *Customer service coordinators* who find efficient recovery solutions for passengers and coordinate with airline operations controllers to provide an assessment of the impact on passengers of possible operations decisions.
- *Dispatchers* who provide flight plans and relevant information to pilots.
- An *air traffic control group* that collects and provides information, such as the likelihood of future ground delay programs, to airline operations controllers.

The AOCC is complemented by *Station Operations Control Units*, located at airport stations, responsible for *local* decisions, such as the assignment of flights to gates, ground workforce to aircraft, and personnel for passenger service.

Airline operations recovery is replete with challenges, including:

1. The recovery solution must take into account the recent flying history of the aircraft, crews, and passengers to ensure that crew work rules are satisfied, aircraft maintenance and safety regulations are met, passengers

are transported to their desired destinations, and aircraft are positioned appropriately at the end of the recovery period.
2. The recovery solution can utilize additional resources, namely reserve crews and spare aircraft (Sohoni et al., 2002, 2003).
3. There are multiple recovery objectives, namely: minimizing the cost of reserve crews and spare aircraft used; minimizing passenger recovery costs; minimizing the amount of time to resume the original plan; and minimizing loss of passenger goodwill.
4. The recovery problem often must be solved quickly, often within minutes.

To meet these challenges, most airline recovery processes are sequential (Rosenberger et al., 2003). The first step in the process is to recover aircraft, with decisions involving flight leg cancellation or delay, and/or aircraft re-routing. The second step is to determine crew recovery plans, assigning crews to uncovered flight legs by reassigning them or utilizing reserve crews. Finally, the third step is for customer service coordinators to develop passenger re-accommodation plans for *disrupted passengers*. A disrupted passenger is one whose planned itinerary is *broken* and impossible to execute during operations because: (1) at least one of the flight legs in the itinerary is canceled; or (2) the connection time between consecutive flight legs in the itinerary is too short due to flight delays. While the AOCC decision process is hierarchical in nature, airline operations controllers, crew planners, and passenger service coordinators consult with one another during the process to assess the feasibility and impact of possible decisions.

This sequential decision process, first aircraft, then crew, and finally passenger recovery, is reflected in the research on airline recovery performed to date. In the following subsections, we present selected airline recovery research.

6.1 Aircraft recovery

When schedule disruptions occur, the aircraft recovery problem is to determine flight departure times and cancellations, and revised routings for affected aircraft. Re-routing options include ferrying (repositioning an aircraft without passengers to another location, where it can be utilized); diverting (flying to an alternate airport); over-flying (flying to another scheduled destination); and swapping (flight legs are re-assigned among different aircraft). These modifications must satisfy maintenance requirements, station departure curfew restrictions and aircraft balance requirements, especially at the start and end of the recovery period. At the end of the period, aircraft types should be positioned to resume operations as planned.

The aircraft balance requirements add complexity to cancellation decisions. Normally, to ensure that aircraft are positioned where needed to fly downstream flight legs, cancellations involve cycles of 2 or more flight legs. To cancel only a single flight leg l and still be able to execute the remaining schedule, it is necessary to deploy a spare aircraft of the type assigned to the destination of

leg l. Because spare aircraft are typically in very limited supply, canceling only a single flight leg is not usually an option.

Beyond the inherent complexities of re-routing aircraft, scheduling delayed flight departures and making cancellation decisions, an effective aircraft recovery solution approach accounts for the downstream costs and impacts on crew and passengers. The extent to which these complexities are captured in models varies, with increasing sophistication achieved over time.

Arguello et al. (1997) present an integrated aircraft delays and cancellations model and generate sequentially, for each fleet type, a set of aircraft routes that minimize delays, cancellations, and re-routing costs. Their model ensures aircraft balance by matching aircraft assignments with the actual aircraft locations at the beginning of the recovery period and with the planned aircraft locations at the end of the period (that is, the end of the day).

The model includes two types of binary decision variables; namely, maintenance-feasible aircraft routes and schedules, and flight cancellation decisions. An aircraft route is a sequence of flight legs spanning the recovery period, with the origin of a flight leg the same as the destination of its predecessor in the sequence, and the elapsed time between two successive legs at least as great as the minimum aircraft turn time. Routes for aircraft with planned maintenance within the recovery period are not altered to ensure that the modified routes satisfy maintenance requirements.

Let P be the set of aircraft routes, Q be the set of aircraft, and F be the set of flight legs. Aircraft route variable x_j^k equals 1 if aircraft k is assigned to route j and 0 otherwise. Its cost, denoted d_j^k, equals the sum of the delay costs associated with flight delays implied by assigning aircraft k to route j. Note that d_j^k is infinite for each aircraft route j commencing at an airport location other than that of aircraft k at the start of the recovery period. A flight cancellation variable, denoted y_i, is set to 1 if flight leg i is canceled and 0 otherwise. The approximate cost associated with the cancellation of each flight leg i is c_i; h_t equals the number of aircraft needed at airport location t at the end of the recovery period to ensure that the next-day plan can be executed; δ_{ij} is equal to 1 if flight leg i is covered by route j; and b_{tj} is equal to 1 if route j ends at the airport t.

The aircraft recovery model is

$$\min \sum_{k\in Q}\sum_{j\in P} d_j^k x_j^k + \sum_{i\in F} c_i y_i$$

subject to

$$\sum_{k\in Q}\sum_{j\in P} \delta_{ij} x_j^k + y_i = 1 \quad \text{for all flight legs } i, \tag{1}$$

$$\sum_{k\in Q}\sum_{j\in P} b_{tj} x_j^k = h_t$$

for all airports t at the end of the recovery period, (2)

$$\sum_{j \in P} x_j^k = 1 \quad \text{for all aircraft } k, \tag{3}$$

$$x_j^k \in \{0, 1\} \quad \text{for all routes } j \text{ and aircraft } k, \tag{4}$$

$$y_i \in \{0, 1\} \quad \text{for all flights } i \text{ in } F. \tag{5}$$

Constraints (1), together with constraints (4) and (5), require each flight to be included in an assigned route or to be canceled. Constraints (2) ensure that at the end of the day (that is, at the end of the recovery period), aircraft are repositioned so that the plan can be resumed at the start of the next day. Finally, constraints (3) enforce the requirement that each aircraft be assigned to exactly one route, commencing at its location at the start of the recovery period. The objective is to minimize flight cancellation and delay costs.

A challenge in formulating this model is to estimate the objective function costs. Because passengers and crews often travel on more than one aircraft route, the costs of delays and cancellations cannot be expressed exactly as a function of a single flight-leg, or as a function of a single aircraft routing. Instead, these costs depend on the *pairs* or *subsets* of flight legs comprising the passenger and crew connections. Hence, approximate delay and cancellation costs are used in the model.

The Arguello, Bard, and Yu model and heuristic solution approach is applied to a relatively small data set representing the Continental Airlines flight schedule for Boeing 757 aircraft, with 42 flights, 16 aircraft, and 13 airport locations. They report that for over 90% of the instances tested, their approach produces a solution within 10% of the lower bound within 10 CPU seconds.

Rosenberger et al. (2002) extend the Arguello, Bard, and Yu model to include supplementary *slot constraints*. Let A equal the set of allocated arrival slots, $R^k(a)$ equal the set of routes for aircraft k that include legs landing in arrival slot a, and let $|H(j, a)|$ represent the number of flight legs in route j using slot a. Then, the slot constraints are of the form:

$$\sum_{j \in R^k(a)} \sum_{k \in Q} \left|H(j, a)\right| x_j^k \leqslant \alpha_a \quad \text{for all } a \text{ in } A. \tag{6}$$

These constraints ensure that the number of aircraft arriving in each allocated time slot in the recovery period does not exceed the airport's restricted capacity, as mandated by ground delay programs, described in Section 4. Additional work on recovering airline operations under conditions of insufficient airport capacity are reported in Vasquez-Marquez (1991), Richetta and Odoni (1993), Hoffman (1997), Luo and Yu (1997), Andreatta et al. (2000), Carlson (2000), Chang et al. (2001), and Metron Inc. (2001).

The body of literature on aircraft recovery is growing as information technology capabilities expand. Selected additional references include Teodorovic and Guberinic (1984), Teodorovic and Stojkovic (1990), Jarrah et al. (1993), Teodorovic and Stojkovic (1995), Yu (1995), Mathaisel (1996), Rakshit et al. (1996), Talluri (1996), Yan and Yang (1996), Yan and Young (1996), Cao and

Kanafani (1997), Clarke (1997), Lettovsky (1997), Yan and Lin (1997), Yan and Tu (1997), and Thengvall et al. (2000).

6.2 Expanded aircraft recovery

Bratu and Barnhart (2005) analyze the operations of a major US airline for the months of July and August 2000, and report that:

(a) Flight delays are not indicative of the magnitude of delay experienced by disrupted passengers. On the same day that disrupted passengers experienced average delays of 419 minutes, the average delay of nondisrupted passengers was only 14 minutes, nearly matching the average flight delay that day.

(b) Disrupted passenger delays and associated costs are significant. Bratu and Barnhart estimate for the airline they study that disrupted passengers represent just about 4% of passengers but account for more than 50% of the total passenger delay. Associated with these disrupted passengers are direct and indirect costs, which can include lodging, meals, re-booking (possibly on other airlines), and loss of passenger goodwill.

Bratu and Barnhart conclude that delay cost estimates that do not take into consideration the costs of disruption cannot be accurate. Recognizing this, Rosenberger et al. (2003) expand their aircraft recovery model to identify disrupted crews and passengers, and their associated costs, by adding constraints and variables to:

(i) compute the delay of each flight leg that is not canceled;
(ii) determine if a connection is disrupted; and
(iii) identify disrupted crews and passengers.

They then estimate delay costs, separately, for disrupted passengers and crews, and for nondisrupted passengers and crews. These, in turn, are included in the objective function of their extended model to achieve a more accurate estimate of delay costs.

To solve their model, Rosenberger, Johnson, and Nemhauser limit the number of aircraft routes considered using an aircraft selection heuristic in which routes are generated only for a selected subset of aircraft. They evaluate their approach using a stochastic model (Rosenberger et al., 2002) to simulate 500 days of airline operations. Simulated disruptions include two-day unscheduled maintenance delays and severe weather disruptions at hub airports. They compare the results of their extended model that accounts for crew and passenger disruptions with those of the simplified model. They report that, compared to the simplified model's solutions, those generated with the extended model exhibit significant reductions in passenger inconvenience and disruptions, at the expense of on-time schedule performance degradation, increased overall delay, and increased incidence of flight cancellation.

Bratu and Barnhart (2006) report similar findings. They also consider disrupted passengers and crews, and develop an aircraft recovery model to determine flight departure times and cancellations that minimize recovery costs,

including the costs of re-accommodating disrupted passengers and crews, re-routing aircraft, and canceling flight legs. Unlike many of the more recent models, their aircraft routing decision variables are flight-leg based, rather than route-based. This reduces the number of decision variables significantly, allowing them to generate recovery solutions for aircraft, crew, and passengers simultaneously. To ensure the satisfaction of maintenance requirements, they do not allow modification of routes for *maintenance-critical* aircraft, that is, aircraft for which maintenance is scheduled that day. They apply their approach to problem instances containing 303 aircraft, 74 airport locations (3 of which are hubs), 1088 flight legs per day on average, and 307,675 passenger itineraries. They achieve solutions within 30 CPU seconds on a PC and report expected reductions of more than 40% in the number of disrupted passengers, more than 45% in the number of passengers required to overnight at a destination other than that planned, and more than 33% in the total delay minutes of disrupted passengers. To achieve this, total delay minutes of nondisrupted passengers increased by 3.7% and the airline's on-time performance, as measured by the US DOT 15-minute on-time performance metric, worsened. This is an expected result when one considers that intentionally delaying aircraft that otherwise would be on-schedule can reduce passenger misconnections and hence, reduce overall passenger delays.

6.3 Crew recovery

Although aircraft recovery decisions repair broken aircraft schedules, they often result in the disruption of crew schedules. Flight cancellations, delays, diversion, and swap decisions, together with crew illness, all result in the unavailability of crews at the locations needed.

Crew recovery options include *deadheading* of crews (i.e., repositioning crews by flying them as passengers) from their point of disruption to the location of a later flight leg to which they are assigned. Once repositioned, the crews can then resume their original work schedule. Another option is to assign a *reserve crew* to cover the flight legs left *unassigned* by the crew disruption. Reserve crews are back-up crews, not originally assigned to the flight schedules, but pre-positioned at certain locations and available to report to duty, if needed. They are guaranteed a minimum monthly salary, whether or not they are called into work, and they are limited to a maximum number of flying hours per month. In addition to possibly incurring additional reserve crew costs when using reserve crews, airlines usually must also pay the *replaced* crew the *entire* amount originally planned, even if the work was not performed. A third recovery option is to reassign a crew from its original schedule to an alternative schedule. In this case, the new assignment must satisfy all collective bargaining agreements and work rule regulations, including maximum crew work time, minimum rest time, maximum flying time, maximum time-away-from-home, etc. When reassigned, crews are typically paid the *maximum* of the pay asso-

ciated with the original schedule or with the new schedule to which they are assigned.

The crew recovery problem then is to construct new schedules for disrupted and reserve crews to achieve coverage of all flights at minimum cost. Because crew costs constitute a significant portion of airline operating costs, second only to fuel costs, crew planning has garnered significant attention. Crew recovery, however, has received much less attention. One reason is that the crew recovery problem is significantly more difficult. First, because of the time horizon associated with recovery operations, recovery solutions must be generated quickly, in minutes instead of the hours or weeks allowed for planning problems. Moreover, information pertaining to the location and recent flying history of each crew member must be known at all times in order to generate recovery plans for the crew that satisfy the myriad of crew rules and collective bargaining agreements. Finally, the objective function of the crew recovery problem is multidimensional. Researchers often cast the crew recovery objective as a blend of minimizing the incremental crew costs to operate the modified schedule, while returning to the plan as quickly as possible and minimizing the number of crew schedule changes made to do so. By limiting the number of crews affected, the quality of the original crew plans will be preserved to the greatest extent possible. Moreover, returning to plan as quickly as possible helps to avoid further downstream disruptions to aircraft, crew, and passengers.

Due to these challenges, the crew recovery literature is relatively limited. Although both cabin and cockpit crews are disrupted and must be recovered, most recovery research focuses on cockpit crews, who are both more costly and more constrained than cabin crews. Pilots have fewer recovery options because they are qualified to fly only aircraft types with the same crew qualifications.

Yu et al. (2003) focus on cockpit crews and present a crew recovery model and solution approach. They consider a set of aircraft types with the same crew qualifications and a set of crews who (i) are qualified to fly these aircraft; *and* (ii) are disrupted or are candidates who are likely to improve the crew recovery solution through *swaps*. For each of these crews, they construct a set of *feasible* pairings, each beginning at the crew's current location and commencing at or later than the time at which the crew is available. Moreover, the generated pairings satisfy all work rules and regulations, considering the amount of work completed by the crew up to the point of disruption. Pairings selected in the recovery solution satisfy *cover constraints* ensuring that each flight leg is either canceled or assigned to one or more crews. When more than one crew is assigned, the additional crews are deadheaded and repositioned to their home location or to another location where they can resume work. The objective is to minimize the sum of (1) deadheading costs; (2) modified crew schedule costs; and (3) cancellation costs due to leaving flight legs uncovered.

Yu et al. define the following sets and parameters:

e equipment type (consisting of one or more crew compatible aircraft types)

I set of active flights to be covered by crews of equipment type e
K set of active and reserve crews available for equipment type e
J_k set of potential feasible pairings for crew k
c_j^k cost of assigning crew k to pairing j
u_i cost of not covering flight leg i
q_k cost of not assigning a pairing to crew k
d_i cost of each crew deadheading on flight i
a_{ij} equal to 1 if flight leg i is included in pairing j; and 0 otherwise.

The variables are:

x_j^k equal to 1 if crew k is assigned to pairing j; and 0 otherwise
z_k equal to 1 if crew k has no pairing assigned; and 0 otherwise
y_i equal to 1 if flight leg i is not covered (is canceled); and 0 otherwise
s_i equal to the number of crews deadheading on flight leg i.

They then formulate the crew recovery problem as

$$\min \sum_{k\in K}\sum_{j\in J_k} c_j^k x_j^k + \sum_{i\in I} u_i y_i + \sum_{k\in K} q_k z_k + \sum_{i\in I} d_i s_i$$

subject to

$$\sum_{k\in Q}\sum_{j\in J_k} a_{ij} x_j^k + y_i - s_i = 1 \quad \text{for all } i \in I, \tag{7}$$

$$\sum_{j\in J_k} x_j^k + z_k = 1 \quad \text{for all } k \in K, \tag{8}$$

$$x_j^k \in \{0,1\} \quad \text{for all } k \in K, \text{ all } j \in J_k, \tag{9}$$

$$y_f \in \{0,1\} \quad \text{for all } f \in I, \tag{10}$$

$$z^k \in \{0,1\} \quad \text{for all } k \in K, \tag{11}$$

$$s_f \in \{0,1,2,\ldots\} \quad \text{for all } f \in I. \tag{12}$$

Constraints (7) ensure that all flight legs are canceled or covered at least once, with s_i representing the number of crews deadheading on flight leg i. Constraints (8) determine whether crew k is assigned to a pairing or must be deadheaded to its crew base, that is, its domicile. Integrality of the solution is guaranteed by constraints (9)–(12).

Yu et al. state that, for typical instances, there are millions of potential crew pairings and hence, the size of the crew recovery model renders exact solution approaches impractical. Using a procedure of Wei et al. (1997), they search heuristically for solutions. They modify or generate a few pairings at a time and test the quality of the solution, and then repeat the process if necessary. Using data provided by Continental Airlines, they evaluate their heuristic approach on instances corresponding to disruptions affecting 1–40 flight legs of the airline's Boeing 737 fleet. Within 8 minutes at most, they generate near-optimal solutions, achieving at most an average 5% optimality gap.

Lettovsky (1997) and Lettovsky et al. (2000) present a similar model, but they include additional constraints restricting the number of crews deadheading on each flight leg to the maximum available capacity. Moreover, their flight cancellation costs include costs of re-assigning passengers to other flights, associated hotel, and meal costs, and estimates of the loss of passenger goodwill. They design a heuristic solution approach for their model that keeps intact as many as possible of the crew schedules, altering only those of disrupted crews and of a few additional crews who greatly facilitate the recovery of crew operations. In restricting the set of crews for which new schedules are generated, *optimality* of the original crew schedules is preserved for crews not affected by modifications to the plan, and the size of the problem is contained, improving tractability and allowing quicker solution times. Lettovsky and Lettovsky, Johnson and Nemhauser describe heuristics to select the crews whose schedules might be altered.

Stojkovic et al. (1998) also address the operational crew scheduling problem, and present a set partitioning model and a branch-and-price algorithm to determine modified monthly schedules for selected crew members. The objective is to cover all tasks at minimum cost while minimizing the number of changes to the planned crew schedules. They generate test problems from pairings of a US airline and report that quality solutions are obtained in reasonable run times.

6.4 Passenger recovery

Just as aircraft recovery decisions result in crew disruptions, aircraft and crew recovery decisions lead to passenger disruptions. The next step of the recovery process then is to reassign disrupted passengers to alternative itineraries, commencing at the disrupted passenger locations after their *available* times, and terminating at their destination, or a location nearby. Disrupted passengers can be assigned to itineraries beginning at least some minimum connection time after the time of their disruption. Only disrupted passengers can be reassigned, and nondisrupted passengers cannot be displaced by reassigned passengers. Clarke (2005) presents modeling strategies for re-accommodating passengers who are disrupted by operations, or by schedule changes resulting from considerations such as revenue management. Barnhart et al. (2002) cast this problem as a multicommodity network flow problem. They let x_p^r represent the number of disrupted passengers originally scheduled on itinerary p who are re-accommodated on itinerary r. In addition to other itineraries offered by the airline, passengers can be accommodated on itineraries offered by other airlines, or itineraries on different modes of transportation. In fact, an alternate itinerary might be the *null* itinerary representing canceled trips, a valid choice for passengers who are disrupted before departing their origin. The planned arrival time at the destination of itinerary p is $l(p)$, and $a(r)$ represents the actual arrival time at the destination of itinerary r. The set of flight legs is F; d_f is the number of seats available

for disrupted passengers, that is, the total number of seats less the number of seats occupied by nondisrupted passengers, on flight f; δ_f^r equals 1 if flight f is contained in itinerary r, and equals 0 otherwise; and n_p is the total number of disrupted passengers of type p. The passenger re-assignment model is then formulated as

$$\min \sum_{p \in P} \sum_{r \in R(p)_k} \big(a(r) - l(p)\big) x_p^r \tag{13}$$

subject to

$$\sum_{p \in P} \sum_{r \in R(p)_k} \delta_f^r x_p^r \leqslant d_f \quad \text{for all } f \in F, \tag{14}$$

$$\sum_{r \in R(p)} x_p^r = n_p \quad \text{for all } p \in P, \tag{15}$$

$$x_p^r \in \{0, 1\} \quad \text{for all } p \in P, \text{ all } r \in R(p). \tag{16}$$

The objective (13) is to find the disrupted passenger reassignments that minimize the total delay experienced by disrupted passengers. Flight capacity constraints (14) ensure that only seats not occupied by nondisrupted passengers are assigned to disrupted passengers. Constraints (15) and (16) ensure that each disrupted passenger is reassigned, albeit perhaps to the null itinerary.

The passenger re-assignment model can be solved exactly, but its solution time can become prohibitive for real-time operations as the number of disrupted passengers grows, thus causing the need for column generation solution approaches to be employed. Bratu and Barnhart (2005), however, solve this problem using a flexible heuristic, termed the *Passenger Delay Calculator*, that allows passenger recovery policies (such as frequent flyers first, or first-disrupted-first-recovered) to be enforced. Using their Passenger Delay Calculator, Bratu and Barnhart analyze two months of airline operations and recovery data for a major airline, as well as numerous simulated scenarios. They conclude that:

1. Connecting passengers are almost three times more likely to be disrupted than passengers without connections. However, connecting passengers who miss their connections are often re-accommodated on their *best* alternative itineraries, that is, on itineraries that arrive at their destinations at the earliest possible time, given the timing of the disruptions. In comparison, only about one-half of the passengers disrupted by flight leg cancellations are re-accommodated on their best itineraries. This occurs because the number of misconnecting passengers per flight leg is small relative to the number of passengers disrupted by a flight leg cancellation.
2. The inability to re-accommodate disrupted passengers on their best itineraries is exacerbated by high load factors, with average delay for disrupted passengers increasing exponentially with load factor.

3. Alternative metrics measuring schedule performance, namely flight cancellation rates and the percentage of flights delayed by more than 45 minutes, are better indicators of passenger disruptions than the US DOT 15-minute on-time performance metric.

7 Robust airline scheduling

Robust planning attempts to deal with data uncertainty in a planning model. In airline schedule and resource allocation planning, there are two primary sources of uncertainty: passenger demand and schedule execution. Here we only deal with uncertainties in the execution of the planned schedule.

In the traditional stochastic programming approach to robust planning, it is necessary to estimate the probability of each possible outcome. One then minimizes the expected cost of the planning decisions plus the cost of the recovery that takes place as a result of the decision and outcome. Unfortunately, this approach is completely intractable in the case of airline planning at present because one has to deal with literally millions of very low probability events. Moreover, there is no obvious way to aggregate meaningfully these events in a way that would simplify the analysis. One could consider planning for only the worst possible outcome, but this would be far too conservative and costly.

Nevertheless, the basic idea of including the anticipated costs of recovery into the planning model can be very useful. A planner needs to think of an optimal plan as being one for which the combined planned and recovery costs, that is, the realized costs, are minimized. This definition of optimality is at odds with the one typically employed by airline optimizers, who have historically excluded recovery costs and optimized only planned costs. In doing so, resource utilization is maximized, with nonproductive time on the ground, i.e., *slack time*, minimized. Lack of slack, however, makes it difficult for disruption to be absorbed in the schedule and limits the number of options for recovery. One should not think of this omission as an oversight on the part of the airlines. Rather, it is based on recognition of the inherent difficulty of including recovery costs in a planning model.

Enhancing schedule planning models to account for recovery costs presents both modeling and computational challenges. A number of researchers have begun to consider this challenge, recognizing that robust planning is a problem rich in opportunity and potential impact. To facilitate the generation of robust plans, they have developed various proxies of *robustness*, mainly focused on finding *flexible* plans that provide a rich set of recovery options for passengers, crews, and aircraft; or plans that isolate the effects of disruptions, requiring only localized plan adjustments.

In the following subsections, we highlight selected work on robust airline planning, while outlining briefly the various modeling and algorithmic approaches employed.

7.1 Robust schedule design

In this section, we describe some recent work that represents a step towards achieving flight schedule designs that are resilient when it comes to passenger disruption. It extends the cost minimization approaches described in Simpson (1966), Chan (1972), Soumis et al. (1980), Etschmaier and Mathaisel (1985), Berge (1994), Marsten et al. (1996), Erdmann et al. (1999), Armacost et al. (2002), and Lohatepanont and Barnhart (2004).

Lan et al. (2005) develop a new approach to minimize the number of passenger misconnections by re-timing flight departures, while keeping all fleeting and routing decisions fixed. Moving flight leg departure times provides an opportunity to re-allocate slack time to reduce passenger disruptions and maintain aircraft productivity. Levin (1971) proposed the idea of adding time windows to fleet routing and scheduling models. Related research can be found in Desaulniers et al. (1997), Klabjan et al. (2002), Rexing et al. (2000), and Stojkovic et al. (2002).

To illustrate the idea, consider the example in Figure 8. In the planned first-in-first-out aircraft routing, flight leg $f1$ is followed by leg $f2$ in one aircraft's rotation, and leg $f3$ is followed by leg $f4$ in another aircraft's rotation. Assume that $f1$ is typically delayed, as indicated in Figure 8. Because insufficient turn time results from the delay, some of the delay to $f1$ will propagate downstream to $f2$, as shown in Figure 7. Then, assuming that $f3$ is typically on-schedule, expected delays are reduced by changing the planned aircraft rotations to $f1$ followed by $f4$, and $f3$ followed by $f2$, as shown in Figure 8.

Lan, Clarke, and Barnhart apply their re-timing model to the flight schedule of a major US airline and compare passenger delays and disruptions in the original schedules with those expected from the solutions to their re-timing model. They find that a 30-minute time window, allowing each flight leg to depart at most 15 minutes earlier or later than in the original schedule, can result in an expected reduction in passenger delay of 20% and a reduction in the number of passenger misconnections of about 40%; a twenty-minute time window can reduce passenger delays by about 16% and reduce the number of passenger misconnections by over 30%; and finally, a ten-minute time window can reduce passenger delays by roughly 10% and passenger misconnections by 20%.

7.2 Robust fleet assignment

Rosenberger et al. (2004) present a robust fleet assignment approach, building on the work reported in Barnhart et al. (1998a). They identify hub inter-connectivity as an important indicator of schedule robustness. Because schedules are sensitive to disruptions at hubs, a more robust schedule is one in which hubs are *isolated* to the greatest extent possible. They quantify the degree to which a hub is isolated using a *hub connectivity* metric; the smaller the value of hub connectivity, the more isolated the hub.

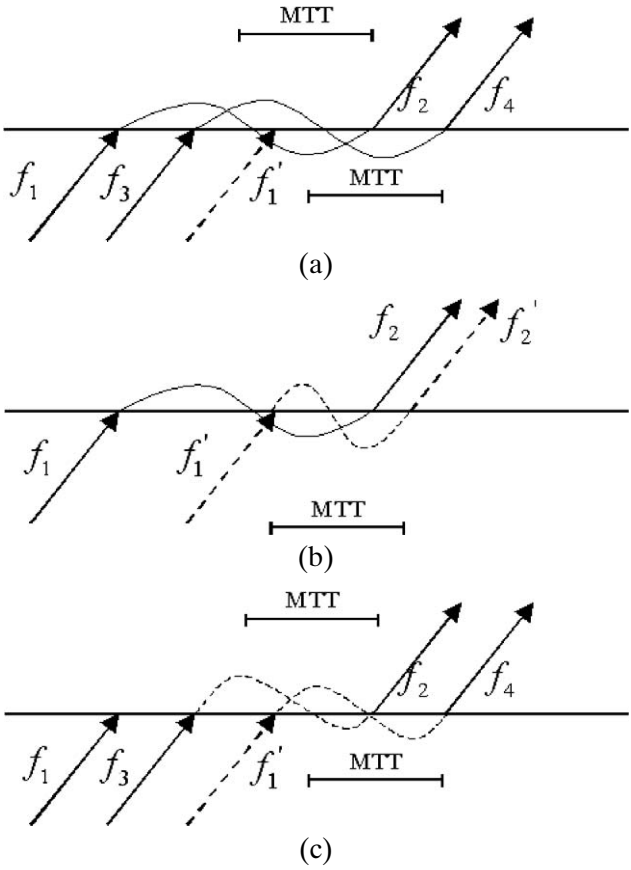


Fig. 8. (a) First-in-first-out routings and delayed flight leg $f1$. (b) Delay propagation due to delay of flight leg $f1$. (c) Revised routings minimizing delay propagation.

Rosenberger, Johnson, and Nemhauser characterize a robust fleet assignment as one with limited total hub connectivity *and* many *short cycles* (cycles with a small number of flight legs). Short cycles allow an airline to limit the number of flights canceled when a cancellation is necessary, thereby lessening the impact of disruptions and facilitating recovery. They let J denote the set of fleets and S be the set of *strings*, or sequences of flight legs beginning at a hub, ending at a hub, and flown by the same aircraft. The hub connectivity metric h_{js} for string $s \in S$ and aircraft of type $j \in J$ equals the number of legs in s if s begins and ends at different hubs, and equals 0 otherwise. The maximum value of hub connectivity is specified by a threshold value ς. For each fleet type $j \in J$ and string $s \in S$, c_{js} is the cost of flying s with aircraft of type j and decision variable x_{js} equals 1 if j is assigned to s, and 0 otherwise. The set of feasible fleet assignment solutions is χ. Their model to determine robust fleet

assignments with hub isolation and short cycles is then

$$\begin{aligned} \min & \sum_{s\in S}\sum_{j\in J} c_{js}x_{js} \\ \text{subject to} & \sum_{s\in S}\sum_{j\in J} h_{js}x_{js} \leqslant \varsigma, \\ & x \in \chi. \end{aligned}$$

The objective is to find the minimum cost fleet assignment with total value of hub connectivity not greater than ς. They propose a related model in which the objective is to maximize hub isolation and limit total fleeting costs to some pre-specified threshold. Using SIMAIR, they compare the solutions to their robust fleet assignment models with those obtained solving a traditional FAM model. They report that with small increases in planned costs it is possible to reduce cancellations significantly and also improve on-time performance.

Smith (2004) focuses on the revenue aspects of robustness in fleet assignment. He adds *purity* to fleet assignment solutions, which means that he limits the number of fleet types at spokes to at most one or two. This, of course, increases planned cost, but purity adds robustness in operations by enhancing the possibility of crew swaps. It also decreases maintenance cost because the need for spare parts is reduced. Interestingly, adding the upper bound constraints on fleet types at spokes makes the FAM model much harder to solve. Smith introduces a station decomposition approach to solve this more difficult FAM model.

7.3 Robust aircraft routing

Maintenance routing, as surveyed in Klabjan (2003), provides an attractive opportunity for adding robustness because modifying routes has a minimal impact on planned cost. Therefore it is not necessary to make an explicit tradeoff between planned cost and robustness.

7.3.1 Degradable airline scheduling

Kang and Clarke (2002) attempt to achieve robustness by isolating the effects of disruptions. They partition the legs of the flight schedule into independent subnetworks, which are determined through alternative models and approaches, each applicable at a particular step of schedule planning, such as schedule design, fleet assignment, or aircraft maintenance routing. The model solutions are constrained to ensure that aircraft (and ultimately, crew) are assigned only to flight legs within a single subnetwork, prohibiting them to operate between subnetworks. (Passengers, on the other hand, can travel within multiple subnetworks.) The subnetworks are prioritized based on the total revenue of the flights legs they contain, with the maximum-revenue subnetwork having the highest priority. When disruptions occur, the top priority subnetworks are recovered first, shielding the associated crew, aircraft, and

passengers to the greatest extent possible from the resulting delays. This has the effect of relegating disruptions to the low priority subnetworks, and minimizing the revenue associated with delayed and disrupted passengers.

An advantage of the approach of Kang and Clarke is that it can simplify recovery. Because delays and propagation effects are contained within a single subnetwork, the recovery process needs only to take corrective action on the flights in the affected subnetwork, and not on the entire airline network.

7.3.2 Robust aircraft routing: Allocating slack to minimize delays

Lan et al. (2005) propose aircraft routing models and algorithms aimed at minimizing delay propagation and passenger delay and disruption. They partition flight leg delays into two categories, namely: propagated delay, that is, delay occurring when the aircraft to be used for a flight leg is delayed on its preceding flight leg; and nonpropagated or independent, delay. Propagated delay is a function of an aircraft's routing, while nonpropagated delay is not. The premise underlying their approach is that propagated delay can be reduced through intelligent aircraft routing, that is, by allocating slack optimally to absorb delay propagation. By reducing propagated delay, they expect to achieve a corresponding reduction in passenger delays.

Using historical delay data, they first estimate the expected independent delay for each flight leg, and then, using these estimates, compute the expected propagated delay for each possible aircraft routing by sequentially computing the earliest departure time for each subsequent flight leg in the routing, given the expected independent delay for that flight leg plus the resulting propagated delay accumulated to that point. Next, for each fleet type, they solve a *daily* model (that is, one that assumes the flight schedule repeats daily) to select aircraft routes that satisfy maintenance requirements while minimizing propagated delay. In their model, they let S be the set of feasible *strings*, where a string is a sequence of *connected* flight legs (that is, the departure station for flight leg f is the same as the arrival station of f's predecessor and the departure time of flight leg f is not earlier than the arrival time plus minimum turn time of flight leg f's predecessor) beginning and ending at a maintenance station and with elapsed time not greater than the maximum time between maintenance checks. The set of daily flight legs is F, F^+ is the set of flight legs originating at a maintenance station, and F^- is the set of flight legs terminating at a maintenance station. The set of ground arcs (including the overnight or wrap-around arcs to ensure that the flight schedule can repeat daily) is denoted by G. The set of strings ending with flight leg i is S_i^-, and the set of strings beginning with flight leg i is S_i^+. They include one binary decision variable x_s for each feasible string s, and ground variables y to count the number of aircraft on the ground at maintenance stations. The delay propagated from flight leg i to flight leg j if flight leg i and flight leg j are in string s is pd_{ij}^s. If flight leg i is in string s, a_{is} equals 1, otherwise it equals 0. Ground variable $y_{i,d}^-$ equals the number of aircraft on the ground before flight leg i departs, and

ground variable $y^+_{i,d}$ equals the number of aircraft on the ground after flight leg i departs. Similarly, ground variable $y^-_{i,a}$ equals the number of aircraft on the ground before flight leg i arrives, and ground variable $y^+_{i,a}$ equals the number of aircraft on the ground after flight leg i arrives. The *count time* is a point in time when aircraft are counted. The number of times string s crosses the *count time* is r_s; p_g is the number of times ground arc g crosses the count time; and N is the number of planes available.

The model to determine robust aircraft routes is

$$\min E\left(\sum_{s\in S}\sum_{(i,j)\in s_k} pd^s_{ij}x_s\right) = E\left(\sum_{s\in S} x_s \sum_{(i,j)\in s_k} pd^s_{ij}\right)$$
$$= \min\left(\sum_{s\in S} x_s E\left(\sum_{(i,j)\in s_k} pd^s_{ij}\right)\right) \tag{17}$$

subject to

$$\sum_{s\in S} a_{is}x_s = 1 \quad \text{for all } i \in F, \tag{18}$$

$$\sum_{s\in Si^+} x_s - y^-_{i,d} + y^+_{i,d} = 0 \quad \text{for all } i \in F^+, \tag{19}$$

$$\sum_{s\in Si^-} x_s - y^-_{i,a} + y^+_{i,a} = 0 \quad \text{for all } i \in F^-, \tag{20}$$

$$\sum_{s\in S} r_s x_s + \sum_{g\in G} p_g y_g \leqslant 0, \tag{21}$$

$$y_g \geqslant 0 \quad \text{for all } g \in G^-, \tag{22}$$

$$x_s \in \{0, 1\} \quad \text{for all } s \in S^-. \tag{23}$$

The objective (17) is to select strings that minimize the expected total propagated delay. Constraints (18) ensure that each flight leg is contained in exactly one string, while constraints (19) and (20) guarantee that the number of aircraft arriving at a location equal the number departing. Constraint (21) ensures that the total number of aircraft in the solution does not exceed the number available. Constraints (22) and (23) guarantee a nonnegative number of aircraft on the ground at all times, and ensure that the number of aircraft assigned to a string is either 0 or 1, respectively. Because each ground variable can be expressed as a sum of binary string variables, the integrality constraints on the ground variables can be relaxed (Hane et al., 1995).

Their solution approach applies the branch-and-price algorithm (Barnhart et al., 1998b), with column generation to enumerate a relevant subset of string variables. They apply their approach to four different networks, each corresponding to a different fleet type operated by a major US network carrier. They compare their robust routing solution with the routing solution generated by the airline and estimate that their solution can yield average reductions of

11% in the number of disrupted passengers, and 44% in total expected propagated delay minutes. They further report that their solution corresponds to an expected improvement of 1.6% in the airline's Department of Transportation (DOT) on-time arrival rate. This is significant because a 1.6% improvement would allow the airline to improve its position in the DOT's airline on-time rankings, which are publicly available and are often cited as an important indicator of airline performance.

7.3.3 Robust routing through swap opportunities

Ageeva and Clarke (2004) use the constraints of the string-based routing model of Lan et al. (2005) but change the objective function to optimize a different robustness criterion. They attempt to build *flexible* aircraft routings with maximal potential for modification during recovery by adding a reward for each opportunity to swap aircraft. Their objective is to maximize the number of swap opportunities in the routing solution. Aircraft swapping is possible when the routings of two aircraft intersect at least twice. To understand how swaps can mitigate the impact of a delayed or unavailable aircraft, consider an example in which aircraft $a1$ is scheduled to depart station s at time $t1$ but is delayed until time $t2$. Further, assume that aircraft $a2$ is scheduled to depart s at time $t3$ (with $t3 > t2 > t1$) but is available for departure at $t1$. Without swapping, the flight legs assigned to aircraft $a1$ experience delays as great as $t2 - t1$, while the flight legs assigned to aircraft $a2$ are not delayed. With swapping, however, none of the flights legs originally assigned to $a1$ or $a2$ is delayed.

Ageeva and Clarke measure the *robustness* of their solutions using an *opportunity index*, defined as the ratio of the number of actual to potential intersecting partial rotations. They report that using their approach, optimal costs are maintained and robustness of the aircraft routing solution, as measured by the opportunity index, are improved up to 35% compared to solutions generated by a basic routing model devoid of robustness considerations.

7.4 Robust crew scheduling

The crew pairing problem, with a focus on minimizing *planned* crew-related cost, has been studied by many researchers. Survey papers on the subject include Yu (1997), Desaulniers et al. (1998), Clarke and Smith (2000), and Barnhart et al. (2003b). The focus of the more recent body of research on *robust* crew scheduling has instead been on minimizing *realized* cost. The underlying motivation stems from the observation that the realized cost associated with a crew pairing solution often differs significantly in practice from the planned cost. Large additional crew costs are incurred, for example, when reserve crews are called in to complete work assigned to disrupted crewmembers no longer able to perform their originally assigned work. The causes might include crewmembers in the wrong location due to one or more flight leg cancellations, or crewmembers reaching the limit on the maximum allowable duty

time before completing their work due to flight delays. The resulting cost increases can be of the order of many millions of dollars for a large airline. To address this issue, researchers, such as Ehrgott and Ryan (2002), Schaefer et al. (2005), Yen and Birge (2000), and Chebalov and Klabjan (2002) have developed approaches to minimize the sum of planned *and* unplanned crew cost.

7.4.1 *Robust crew pairing: The role of crew connections between different aircraft*

Ehrgott and Ryan (2002) propose an approach to balance costs and robustness in generating crew pairing solutions. For each pairing, they compute its value of *nonrobustness* by approximating downstream effects of delays within the pairing. In their approximation, the value of nonrobustness is zero if crews do not change planes, but equals the potential disruptive effects of delays if the plan requires crews to connect between different aircraft. The objective is to minimize the value of nonrobustness, while maintaining the cost of the corresponding crew pairing solution to less than that of the minimum-cost crew pairing solution plus some pre-specified positive value.

Ehrgott and Ryan report that small increases in cost allow considerable robustness gains. In one instance, by increasing costs by less than 1%, they were able to reduce their metric of "nonrobustness" by more than 2 orders of magnitude. Their more robust solutions are characterized by longer ground times between successive flights on different aircraft within a pairing; fewer aircraft changes within pairings; slightly longer duty times; and a slight increase in the number of pairings in the solution.

7.4.2 *Minimizing expected crew costs*

Yen and Birge (2000) and Schaefer et al. (2005) develop approaches that include both planned and unplanned costs in the objective function of the crew model. Yen and Birge develop a stochastic crew scheduling model and corresponding solution approach, while Schaefer et al. solve a deterministic crew pairing problem with an objective to minimize *expected* crew pairing costs. In Schaefer et al., expected costs are approximated for each pairing under the assumptions that there are no interactions between the pairings and recovery is achieved simply by delaying the next flight in the pairing until the crew is available (pushback). Pushback recovery helps to justify the no interactions assumption, but much more sophisticated recovery procedures are used in practice at hubs. Even with these simplifying assumptions, it is still not possible to calculate the realized cost of pairings analytically. Thus the cost of each pairing considered in the optimization is determined by Monte Carlo simulation.

Schaefer et al. denote J as the set of feasible pairings, F the set of flight legs, $\tilde{c}_j$ the expected cost of pairing j, and let a_{ij} equal 1 if flight leg i is covered by pairing j. The decision variable x_j, for all $j \in P$, equals 1 is pairing j is included in the solution, and equals 0 otherwise.

The robust crew model of Schaefer et al. is

$$\min \sum_{j \in J} \tilde{c}_j x_j \tag{24}$$

subject to

$$\sum_{j \in J} a_{ij} x_j = 1 \quad \text{for all } i \in F^-, \tag{25}$$

$$x_j \in \{0, 1\} \quad \text{for all } j \in P^-. \tag{26}$$

The objective (24) is to minimize the expected crew costs associated with the pairings in the solution. The set of selected pairings must contain each flight leg exactly once (25).

Crew schedules obtained with these expected pairing costs are compared with those obtained by using the standard deterministic costs and also a set of penalty costs whereby attributes of pairings that might lead to poor operational performance are penalized. The attributes considered are sit times between flights when a crew changes planes, flying and elapsed times of duties, and rest time between duties. The operational performance of crew schedules are evaluated using SIMAIR, with only mild disruptions considered, that is, individual delays rather than major disruptions such as those which reduce airport capacity. The standard cost measure of *FTC*, which is total cost in minutes of pay minus minutes of flying time divided by flying time, is used to compare schedules. For the fleets considered from a major airline, the planned FTCs were typically in the 2–4 % range and increased only slightly when either the expected costs or penalty costs were used. The operational FTCs ranged from 4% to 9% and were lowest for the expected cost method and very close to a lower bound. However, in an absolute sense the expected cost solutions performed only marginally better than the deterministic solutions perhaps because of the assumption of mild disruptions only. Note that more severe disruptions would have invalidated the pushback assumption.

7.4.3 Move-up crews

In an approach analogous to the idea in Ageeva and Clarke (2004) of providing flexibility through aircraft swapping, Chebalov and Klabjan (2002) develop the concept of enhancing recovery flexibility through *move-up crews*. A move-up crew for flight i is a crew, not actually assigned to i, but capable of being assigned to i, if necessary. For feasibility of this potential assignment, the move-up crew must have the same crew base, or domicile, as the crew currently assigned to i, must be ready to operate i before the departure time of i, and must end the pairing on the same day as the pairing currently covering i. Chebalov and Klabjan consider move-up crews only for flights i that depart hub locations and do not begin a pairing. The objective is to maximize the number of move-up crews so that recovery is more likely to be effected by swapping the assigned pairings of the delayed crew and an available, alternative crew.

Let J represent the set of feasible pairings; F be the set of flight legs; a_{ij} equal 1 if flight leg i is covered by pairing j; HL designate the set of hub locations; CB be the set of crew base locations; D be the set of the possible number of days remaining in a pairing; $J_{cb,d}$ represent the set of pairings starting at the crew base cb with d days remaining after flight leg i to the end of the pairing; J_i be the set of pairings whose first leg is i; $\overline{J}_{i,cb,d}$ be the set of pairings that yields a move-up crew for flight i covered by a crew originating at cb with d days remaining from flight leg i to the end of the pairing; r be the robustness factor denoting the maximum allowable percentage increase in the cost of the solution to allow increased robustness; and M be an arbitrary number (usually 2 or 3). Chebalov and Klabjan solve the standard crew pairing problem minimizing operating costs with constraints (25) and (26) to obtain the minimum planned crew pairing costs, c_{opt}. They then include in their model four sets of decision variables. If pairing j is included in the solution, x_j, for all $j \in P$, is equal to 1, otherwise it equals 0. If flight leg i is covered by a pairing starting at the crew base cb with d days remaining after flight leg i to the end of the pairing, $y_{i,d}^{cb}$ is equal to 1, otherwise it equals 0. If flight leg i is covered by a pairing whose first leg is i, w_i is equal to 1, otherwise it equals 0. The number of move-up crews for flight leg i, if i is a leg originating at a hub $h \in HL$, is denoted by $z_{i,d}^{cb}$.

The Chebalov and Klabjan model is

$$\max \sum_{i \in F} \sum_{cb \in CB} \sum_{d \in D} z_{i,d}^{cb} \tag{27}$$

subject to

$$\sum_{j \in J_{cb,d}} a_{ij} x_j = y_{i,d}^{cb} \quad \text{for all } i \in F \text{ originating at any hub,} \tag{28}$$

$$\sum_{j \in J} a_{ij} x_j = 1 \quad \text{for all } i \in F \text{ originating at any spoke,} \tag{29}$$

$$\sum_{j \in J_i} x_j = w_i \quad \text{for all } i \in F \text{ originating at any crew base,} \tag{30}$$

$$w_i + \sum_{cb \in CB} \sum_{d \in D} y_{i,d}^{cb} = 1$$
$$\text{for all } i \in F \text{ originating at any crew base,} \tag{31}$$

$$\sum_{cb \in CB} \sum_{d \in D} y_{i,d}^{cb} = 1$$
$$\text{for all } i \in F \text{ originating at a hub but not a crew base,} \tag{32}$$

$$\sum_{j \in \overline{J}_{i,cb,d}} x_j \geqslant z_{i,d}^{cb} \quad \text{for all } i \in F \text{ originating at any hub,} \tag{33}$$

$$z_{i,d}^{cb} \leqslant M y_{i,d}^{cb} \quad \text{for all } i \in F \text{ originating at any hub,} \tag{34}$$

$$\sum_j c_j x_j \leqslant (1+r)c_{\text{opt}}, \tag{35}$$

$$x_j \in \{0,1\} \quad \text{for all } j \in J, \tag{36}$$

$$w_i \in \{0,1\} \quad \text{for all } i \in F, \tag{37}$$

$$y_{i,d}^{cb} \in \{0,1\} \quad \text{for all } i \in F, d \in D, cb \in CB, \tag{38}$$

$$z_{i,d}^{cb} \in \{0,1,\ldots,M\} \quad \text{for all } i \in F, d \in D, cb \in CB. \tag{39}$$

Constraints (29) and (36) ensure that each flight leg i is covered by exactly one pairing. With constraints (37)–(39), constraints (30) identify flight legs that are the first leg in a pairing and constraints (28), (31), and (32) identify flight legs that are candidates for move-up crews. The number of move-up crews for flight leg i is bounded by 0 (constraints (33)) if i originates at a spoke location or is the first leg in a pairing, and otherwise it is bounded by the minimum of M (usually 2 or 3), and the number of eligible move-up crews for i (constraints (34)). Constraint (35) ensures that the pairing solution has cost no greater than an allowable tolerance above the minimum cost pairing. The objective (27) is to maximize the total number of move-up crews for all flight legs.

Chebalov and Klabjan present a Lagrangian decomposition method to solve their model. They perform computational experiments and report, that for certain instances, there are crew solutions characterized by only slightly higher planned costs and by a 5- to 10-fold increase in the number of move-up crews, compared to the optimal solutions to the more conventional, cost-minimizing crew pairing problem.

8 Conclusions

Flight and crew schedules and passenger itineraries have become increasingly "fragile" as a result of the growing complexity of the air transportation system and the tight coupling of its various elements. The resulting direct and indirect economic costs are very large, certainly amounting to several billion dollars annually. The airline industry has a vital stake in research aimed at mitigating the effects of severe weather and other disruptive events and at expediting recovery from "irregular" operations.

As this chapter has indicated, a very significant body of recent and ongoing work has led to major progress toward these objectives. Two breakthrough developments have been the primary drivers behind this progress. First, Collaborative decision making has made it possible to apply the principles of information sharing and distributed decision making to ATFM, by expanding the databases available to airline and FAA (and, soon, European) traffic flow managers, creating common situational awareness, and introducing shared real-time tools and procedures. And second, there is growing recognition in the airline industry of the fact that planning for schedule robustness

and reliability may be just as important as planning for minimizing costs in the complex, highly stochastic and dynamic environment of air transportation. Specific achievements that have been described herein include: improved understanding and better modeling of the physics of airport and airspace capacity and delays (Section 2); realization of the need for market-based mechanisms to supplement widely-used administrative methods for allocating scarce airport capacity among prospective airport users (Section 3); development of models and optimization tools to support GDP decision-making under a wide range of conditions, including the presence of uncertainty regarding forecast capacity and demand (Section 4); development and implementation of airline-based models for efficient "recovery" of aircraft, crews, and passengers following schedule disruptions (Section 6); and the nascent appearance of increasingly viable models for introducing robustness in airline route design and in the scheduling of aircraft, crews, and passengers (Section 7).

At the same time, it is fair to describe all this work as still being in its early stages in many respects – an assessment that applies equally well to the domains of both the airlines and the providers of air traffic management services. For example, in the case of the latter, approaches for dealing with uncertainty – an altogether critical issue in the ATM and ATFM context – are still quite far removed from being applied in practice. Integrated consideration and optimization of both arrival *and* departure schedules at GDP airports could also offer significant improvements over the existing approaches that focus solely on arrivals. Research on collaborative routing is still in its infancy. On the side of the airlines, decision support software for recovery is perhaps at the stage where planning software was 15 years ago. While research is active and hardware and data support have improved substantially, optimization-based decision support tools for rapid recovery are still at an early stage of implementation at the major airlines. Finally, and most important, a full integration of the emerging ATFM CDM philosophy and associated models with airline recovery planning models and robust scheduling models has not even begun. This represents a difficult, but crucial future research challenge.

Acknowledgements

The work of the first and fourth authors was supported in part by NEXTOR, the National Center for Excellence in Aviation Operations Research, under Federal Aviation Administration cooperative agreement number 01CUMD1. The work of the second and fourth authors was supported in part by the Alfred P. Sloan Foundation as part of the MIT Global Airline Industry Program.

References

Ageeva, Y., Clarke, J.P. (2004). Incorporating robustness into airline scheduling. Working paper, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA.

Andreatta, G., Odoni, A.R. (2003). Analysis of market-based demand management strategies. In: Ciriani, T., Fasano, G., Gliozzi, S., Tadei, R. (Eds.), *Operations Research in Space and Air*. Kluwer Academic, pp. 257–278.

Andreatta, G., Brunetta, L., Guastalla, G. (2000). From ground holding to free flight: An exact approach. *Transportation Science* 34, 394–401.

Arguello, M.F., Bard, J.F., Yu, G. (1997). Models and methods for managing airline irregular operations. In: Yu, G. (Ed.), *Operations Research in the Airline Industry*. Kluwer Academic, Boston, MA, pp. 1–45.

Armacost, A., Barnhart, C., Ware, K. (2002). Composite variable formulations for express shipment service network design. *Transportation Science* 36, 1–20.

ATAC Corporation (2003). Airport and airspace simulation model. Available at *http://www.atac.com/prodsvs/simmod.htm*.

Balinski, M.L., Shihidi, N. (1998). A simple approach to the product rate variation problem via axiomatics. *Operations Research Letters* 22, 129–135.

Ball, M.O., Lulli, G. (2004). Ground delay programs: Optimizing over the included flight set based on distance. *Air Traffic Control Quarterly* 12, 1–25.

Ball, M., Dahl, R., Stone, L., Thompson, T. (1993). OPTIFLOW build-I design document, version 1.5. Optiflow project report to the FAA, December.

Ball, M.O., Chen, C., Hoffman, R., Vossen, T. (2001a). Collaborative decision making in air traffic management: Current and future research directions. In: Bianco, L., Dell'Olmo, P., Odoni, A. (Eds.), *New Concepts and Methods in Air Traffic Management*. Springer-Verlag, Berlin, pp. 17–30. A preprint appeared in *Proceedings of ATM'99, Workshop on Advanced Technologies and Their Impact on Air Traffic Management in the 21st Century*, 1999.

Ball, M.O., Hoffman, R., Knorr, D., Wetherly, J., Wambsganss, M. (2001b). Assessing the benefits of collaborative decision making in air traffic management. In: Donohue, G., Zellweger, A. (Eds.), *Air Transportation Systems Engineering*. American Institute of Aeronautics and Astronautics, Inc., Reston, VA, pp. 239–250. A preprint appeared in *Proceedings of 3rd USA/Europe air Traffic Management R&D Seminar*, 2000.

Ball, M.O., Vossen, T., Hoffman, R. (2001c). Analysis of demand uncertainty in ground delay programs. In: *Proceedings of 4th USA/Europe Air Traffic Management R&D Seminar*, Santa Fe, NM.

Ball, M.O., Hoffman, R., Odoni, A., Rifkin, R. (2003). Efficient solution of a stochastic ground holding problem. *Operations Research* 51, 167–171.

Ball, M.O., Donohue, G., Hoffman, K. (2005). Auctions for the safe, efficient and equitable allocation of airspace system resources. In: Cramton, P., Shoham, Y., Steinberg, R. (Eds.), *Combinatorial Auctions*. MIT Press, Cambridge, pp. 507–538.

Barnhart, C., Boland, N., Clarke, L., Johnson, E., Nemhauser, G., Shenoi, R. (1998a). Flight string models for aircraft fleeting and routing. *Transportation Science* 32, 208–220.

Barnhart, C., Johnson, E., Nemhauser, G., Savelsbergh, M., Vance, P. (1998b). Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46, 316–329.

Barnhart, C., Kniker, T., Lohatepanont, M. (2002). Itinerary-based airline fleet assignment. *Transportation Science* 36, 199–217.

Barnhart, C., Belobaba, P., Odoni, A.R. (2003a). Applications of operations research in the air transport industry. *Transportation Science* 37, 368–391.

Barnhart, C., Cohn, A.M., Johnson, E.L., Klabjan, D., Nemhauser, G.L., Vance, P.H. (2003b). Airline crew scheduling. In: Hall, R.W. (Ed.), *Handbook of Transportation Science*, 2nd edition. Kluwer Academic, Norwell, MA, pp. 517–560.

Beasley, J., Sonander, J., Havelock, P. (2001). Scheduling aircraft landings at London Heathrow using a population heuristic. *Journal of the Operational Research Society* 52, 483–493.

Beatty, R., Hsu, R., Berry, L., Rome J. (1998). Preliminary evaluation of flight delay propagation through an airline schedule. In: *Proceedings of the Second USA/Europe Air Traffic Management R&D Seminar*, Orlando, Fl.

Berge, M. (1994). Timetable optimization: Formulation, solution approaches, and computational issues. In: *AGIFORS Proceedings*, pp. 341–357.

Berge, M., Hopperstad, C., Heraldsdottir, A. (2003). Airline schedule recovery in collaborative flow management with airport and airspace capacity constraints. In: *Proceedings of the Fifth USA/Europe Air Traffic Management R&D Seminar*, Budapest, Hungary.

Bertsimas, D., Stock Paterson, S. (1998). The air traffic flow management problem with en route capacities. *Operations Research* 46, 406–422.

Bertsimas, D., Stock Paterson, S. (2000). The traffic flow management rerouting problem in air traffic control: A dynamic network flow approach. *Transportation Science* 34, 239–255.

Bhogadi, N. (2002). Modeling demand uncertainties during ground delay programs. MSSE thesis, University of Maryland.

Bilimoria, K., Sridhar, B., Chatterji, G., Sheth, K., Grabbe, S. (2000). FACET: Future ATM concepts evaluation tool. In: *Proceedings of 3rd USA/Europe Air Traffic Management R&D Seminar*, Napoli, Italy.

Blumstein, A. (1959). The landing capacity of a runway. *Operations Research* 7, 752–763.

Bratu, S. (2003). Airline customer service network design and operations. PhD thesis, Department of Aeronautics and Astronautics and the Center for Transportation and Logistics, Massachusetts Institute of Technology, Cambridge, MA.

Bratu, S., Barnhart, C. (2005). An analysis of passenger delays using flight operations and passenger booking data. *Air Traffic Control Quarterly* 13, 1–27.

Bratu, S., Barnhart, C. (2006). Flight operations recovery: New approaches considering passenger recovery. *Journal of Scheduling* 9 (3), 279–298.

Bruckner, J.K. (2002). Airport congestion when carriers have market power. *American Economic Review* 92, 1357–1375.

Cao, J., Kanafani, A. (1997). Real-time decision support for integration of airline flight cancellations and delays, Part I: Mathematical formulation. *Transportation Planning and Technology* 20, 183–199.

Carlin, A., Park, R. (1970). Marginal cost pricing of airport runway capacity. *American Economic Review* 60, 310–318.

Carlson, P.M. (2000). Exploiting the opportunities of collaborative decision making: A model and efficient solution algorithm for airline use. *Transportation Science* 34, 381–393.

Chan, Y. (1972). Route network improvement in air transportation schedule planning. Flight Transportation Laboratory Report R72-3, Massachusetts Institute of Technology, Cambridge, MA.

Chang, K., Howard, K., Oiesen, R., Shisler, L., Tanino, M., Wambsganss, M.C. (2001). Enhancements to the FAA ground-delay program under collaborative decision making. *Interfaces* 3, 57–76.

Chebalov, S., Klabjan, D. (2002). Robust airline crew scheduling: Move-up crews. In: *Proceedings of the, 2002 NSF Design, Service, and Manufacturing Grantees and Research Conference*.

Clarke, J.-P., Melconian, T., Bly, E., Rabbani, F. (2004). MEANS – The MIT extensible air network simulation. *Simulation: Transactions of the Society International for Computer Simulation*, in press.

Clarke, M. (1997). Development of heuristic procedures for flight rescheduling in the aftermath of irregular airline operations. ScD dissertation, Massachusetts Institute of Technology, Cambridge MA.

Clarke, M. (2005). Passenger reaccommodation a higher level of customer service. Presented at *AGIFORS Airline Operations Study Group Meeting*, Mainz, Germany.

Clarke, M., Smith, B. (2000). The impact of operations research on the evolution of the airline industry: A review of the airline planning process. Research paper, Sabre Inc., Dallas, TX.

Clarke, M., Lettovsky, L., Smith, B. (2000). The development of the airline operations control center. In: Butler, G., Keller, M.R. (Eds.), *Handbook of Airline Operations*. Aviation Week Group of McGraw-Hill, Washington, DC.

Daniel, J. (1995). Congestion pricing and capacity at large hub airports: A bottleneck model with stochastic queues. *Econometrica* 63, 327–370.

de Neufville, R., Odoni, A. (2003). *Airport Systems: Planning, Design and Management*. McGraw-Hill, New York.

Dear, R.G., Sherif, Y.S. (1991). An algorithm for computer assisted sequencing and scheduling of terminal area operations. *Transportation Research A* 25, 129–139.

Desaulniers, G., Desrosiers, J., Solomon, M.M., Soumis, F. (1997). Daily aircraft routing and scheduling. *Management Science* 43, 841–855.

Desaulniers, G., Desrosiers, J., Gamache, M., Soumis, F. (1998). Crew scheduling in air transportation. In: Crainic, T., Laporte, G. (Eds.), *Fleet Management and Logistics*. Kluwer Academic, pp. 169–185.

Dobbyn, T. (2000). US prepares plan to cut summer flight delays. *Reuters Wire*, March 3.

DotEcon Ltd. (2001). Auctioning airport slots. Report for HM Treasury and the Department of the Environment, Transport and the Regions, United Kingdom. Available at *http://www.dotecon.com*.

Ehrgott, M., Ryan, D.M. (2002). Constructing robust crew schedules with bicriteria optimization. *Journal for Multi-Criteria Decision Analysis* 11, 139–150.

Erdmann, A., Nolte, A., Noltemeier, A., Schrader, R. (1999). Modeling and solving the airline schedule generation problem. Technical Report 99-351, ZAIK, University of Cologne, Germany.

Erzberger, H. (1995). Design principles and algorithms for automated air traffic management. AGARD Lecture Series 200, Brussels. Available at *http://www.ctas.arc.nasa.gov/*.

Etschmaier, M.M., Mathaisel, D.F.X. (1985). Airline scheduling: An overview. *Transportation Science* 19, 127–138.

EUROCONTROL (2001). CAMACA: The commonly agreed methodology for airside capacity assessment. Brussels. Available at *http://www.eurocontrol.int/camaca/*.

EUROCONTROL (undated). RAMS homepage. Available at *http://www.eurocontrol.fr/projects/rams/mainwin.html*.

Fan, T.P. (2003). Market-based airport demand management – theory, model and applications. PhD dissertation, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA.

Fan, T.P., Odoni, A.R. (2002). A practical perspective on airport demand management. *Air Traffic Control Quarterly* 10, 285–306.

Federal Aviation Administration (FAA) (2001). Airport capacity benchmark report, US Government Printing Office, Washington, DC. Available at *http://www.faa.gov/events/benchmarks/*.

Filar, J., Manyem, P., White, K. (2000). How airlines and airports recover from schedule perturbations: A survey. *Annals of Operations Research* 108, 315–333.

Gilbo, E.P. (1993). Airport capacity: Representation, estimation, optimization. *IEEE Transactions on Control Systems Technology* 1, 144–154.

Gilbo, E., Howard, K. (2000). Collaborative optimization of airport arrival and departure traffic flow management strategies for CDM. In: *Proceedings of 3rd USA/Europe Air Traffic Management R&D Seminar*.

Hall, W. (1999). Information flows and dynamic collaborative decision-making architecture: Increasing the efficiency of terminal area operations. PhD dissertation, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Hane, C.A., Barnhart, C., Johnson, E.L., Marsten, R.E., Nemhauser, G.L., Sigismondi, G. (1995). The fleet assignment problem: Solving a large-scale integer program. *Mathematical Programming* 70, 211–232.

Hansen, M. (2002). Micro-level analysis of airport delay externalities using deterministic queuing models: A case study. *Journal of Air Transport Management* 8, 73–87.

Hinston, J.M., Aigoin, G., Delahaye, D., Hansman, R.J., Puechmorel, S. (2001). Introducing structural considerations into complexity metrics. In: *Proceedings of 4th USA/Europe ATM R&D Seminar*, Santa Fe, NM.

Hoffman, R. (1997). Integer programming models for ground-holding in air traffic flow management. PhD thesis, The National Center of Excellence for Aviation Operations Research, University of Maryland, College Park, MD.

Hoffman, R., Ball, M.O. (2003). A comparison of formulations for the single-airport ground-holding problem with banking constraints. *Operations Research* 48, 578–590.

Inniss, T., Ball, M.O. (2002). Estimating one-parameter airport arrival capacity distributions for air traffic flow management. *Air Traffic Control Quarterly* 12, 223–252.

International Air Transport Association (IATA) (2000). *Worldwide Scheduling Guidelines*, 1st edition. Montreal, Canada.

Jarrah, A., Yu, G., Krishnamurthy, N., Rakshit, A. (1993). A decision support framework for airline flight cancellations and delays. *Transportation Science* 27, 266–280.

Kang, L.S., Clarke, J.P. (2002). Degradable airline scheduling. Working paper, Global Airline Industry Program, Massachusetts Institute of Technology, Cambridge, MA.

Klabjan, D. (2003). Large scale models in the airline industry. Working paper, Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, IL.

Klabjan, D., Johnson, E.L., Nemhauser, G.L., Gelman, E., Ramaswamy, S. (2002). Airline crew scheduling with time windows and plane count constraints. *Transportation Science* 36, 337–348.

Kotnyek, B., Richetta, O. (2004). Equitable models for the stochastic ground holding problem under collaborative decision-making. Working paper, MSIS Department, University of Massachusetts-Boston, Boston, MA.

Lan, S., Clarke, J.P., Barnhart, C. (2005). Planning for robust airline operations: Optimizing aircraft routings and flight departure times to minimize passenger disruptions. *Transportation Science* 40, 15–28.

Lettovsky, L. (1997). Airline operations recovery: An optimization approach. PhD thesis, Georgia Institute of Technology, Atlanta, GA.

Lettovsky, L., Johnson, E.L., Nemhauser, G.L. (2000). Airline crew recovery. *Transportation Science* 34, 337–348.

Levin, A. (1971). Scheduling and fleet routing models for transportation systems. *Transportation Science* 5, 232–255.

Lohatepanont, M., Barnhart, C. (2004). Airline schedule planning: Integrated models and algorithms for schedule design and fleet assignment. *Transportation Science* 38, 19–32.

Long, D., Lee, D., Johnson, J., Gaier, E., Kostiuk, P. (1999). Modeling air traffic management technologies with a queuing network model of the national airspace system. Report NASA/CR-1999-208988, NASA Langley Research Center, Hampton, VA.

Luo, S., Yu, G. (1997). On the airline schedule perturbation problem caused by the ground delay program. *Transportation Science* 31, 298–311.

Manning, C.A., Mills, S.H., Fox, C.M., Mogilka, H.J. (2002). Using air traffic control taskload measures and communications events to predict subjective workload. Report DOT/FAA/AM-02/4, FAA Office of Aerospace Medicine, Washington, DC.

Marsten, R., Subramanian, R., Gibbons, L. (1996). Junior analyst extraordinaire (JANE). In: *AGIFORS Proceedings*, pp. 247–259.

Mathaisel, D.F.X. (1996). Decision support airline system operations control and irregular operations. *Computers & Operations Research* 23, 1083–1098.

Metron Inc. (2001). Collaborative decision making. Available at *http://www.metsci.com/cdm*.

Midkiff, A.H., Hansman, R.J., Reynolds, T.G. (2004). Air carrier flight operations. Report ICAT-2004-3, MIT International Center for Air Transportation, Massachusetts Institute of Technology, Cambridge, MA.

Morrison, S. (1987). The equity and efficiency of runway pricing. *Journal of Public Economics* 34, 45–60.

Nilim, A., El Ghaoui, L., Duong, V., Hansen, M. (2001). Trajectory based air traffic management under weather uncertainty. In: *Proceedings of the Fourth USA/Europe Air Traffic Management R&D Seminar*, Santa Fe, NM.

Niznik, T. (2001). Optimizing the airline response to ground delay programs. Presented at *AGIFORS Symposium*, Ocho Rios, Jamaica.

Odoni, A.R. (1987). The flow management problem in air traffic control. In: Odoni, A.R., Bianco, L., Szego, G. (Eds.), *Flow Control of Congested Network*. Springer-Verlag.

Odoni, A.R., Deyst, J., Feron, E., Hansman, R.J., Khan, K., Kuchar, J.K., Simpson, R. (1997). Existing and required modeling capabilities for evaluating ATM systems and concepts. International Center for Air Transportation, Massachusetts Institute of Technology. Available at *http://web.mit.edu/aeroastro/www/labs/AATT/aatt.html*.

Preston Aviation Solutions (2002). TAAM solutions. Available at *http://www.preston.net/products/TAAM.htm*.

Psaraftis, H.N. (1980). A dynamic programming approach for sequencing groups of identical jobs. *Operations Research* 28, 1347–1359.

Rakshit, A., Krishnamurthy, N., Yu, G. (1996). System operations advisor: A real-time decision support system for managing airline operations at United Airlines. *Interfaces* 26, 50–58.

Rexing, B., Barnhart, C., Kniker, T., Jarrah, A., Krishnamurthy, N. (2000). Airline fleet assignment with time windows. *Transportation Science* 34, 1–20.

Richetta, O., Odoni, A.R. (1993). Solving optimally the static ground-holding policy problem in air traffic control. *Transportation Science* 27, 228–238.

Richetta, O., Odoni, A.R. (1994). Dynamic solutions to the ground-holding problem in air traffic control. *Transportation Research A* 28, 167–185.

Rosenberger, J., Schaefer, A.J., Goldsman, D., Johnson, E.L., Kleywegt, A.J., Nemhauser, G.L. (2002). A stochastic model of airline operations. *Transportation Science* 36, 357–377.

Rosenberger, J., Johnson, E., Nemhauser, G. (2003). Rerouting aircraft for airline recovery. *Transportation Science* 37, 408–421.

Rosenberger, J., Johnson, E.L., Nemhauser, G.L. (2004). A robust fleet-assignment model with hub isolation and short cycles. *Transportation Science* 38, 357–368.

Schaefer, A.J., Johnson, E.L., Kleywegt, A.J., Nemhauser, G.L. (2005). Airline crew scheduling under uncertainty. *Transportation Science* 39, 340–348.

Simpson, R.W. (1966). Computerized schedule construction for an airline transportation system. Report FT-66-3, Flight Transportation Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Smith, B.C. (2004). Robust airline fleet assignment. PhD dissertation, The Logistics Institute, Georgia Institute of Technology, Atlanta, GA.

Sohoni, M.G., Johnson, E.L., Bailey, T.G., Carter, H. (2002). Operational airline reserve crew planning. Working paper, The Logistics Institute, Georgia Institute of Technology, Atlanta GA.

Sohoni, M.G., Johnson, E.L., Bailey, T.G. (2003). Long range reserve crew manpower planning. Working paper, The Logistics Institute, Georgia Institute of Technology, Atlanta GA.

Soumis, F., Ferland, J.A., Rousseau, J.-M. (1980). A model for large scale aircraft routing and scheduling problems. *Transportation Research B* 14, 191–201.

Sridhar, B., Seth, K.S., Grabbe, S. (1998). Airspace complexity and its application in air traffic management. In: *The 2nd USA/Europe ATM R&D Seminar*, Orlando, FL.

Stamatopoulos, M., Zografos, K., Odoni, A.R. (2004). A decision support system for airport strategic planning. *Transportation Research C* 12, 91–118.

Stojkovic, G., Soumis, F., Desrosiers, J., Solomon, M. (2002). An optimization model for a real-time flight scheduling problem. *Transportation Research A* 36, 779–788.

Stojkovic, M., Soumis, F., Desrosiers, J. (1998). The operational airline crew scheduling problem. *Transportation Science* 32, 232–245.

Talluri, K. (1996). Swapping applications in a daily airline fleet assignment. *Transportation Science* 30, 237–284.

Teodorovic, D., Guberinic, S. (1984). Optimal dispatching strategy on an airline network after a schedule perturbation. *European Journal of Operational Research* 15, 178–183.

Teodorovic, D., Stojkovic, G. (1990). Model for operational daily airline scheduling. *Transportation Planning and Technology* 14, 273–286.

Teodorovic, D., Stojkovic, G. (1995). A model to reduce airline schedule disturbances. *Journal of Transportation Engineering ASCE* 121, 324–331.

Terrab, M., Odoni, A.R. (1993). Strategic flow management for air traffic control. *Operations Research* 41, 138–152.

Thengvall, B., Yu, G., Bard, J. (2000). Balancing user preferences for aircraft schedule recovery during irregular airline operations. *IIE Transactions* 32, 181–193.

Vasquez-Marquez, A. (1991). American airlines arrival slot allocation system (ASAS). *Interfaces* 21, 42–61.

Venkatakrishnan, C.S., Barnett, A.I., Odoni, A.R. (1993). Landing time intervals and aircraft sequencing in a major terminal area. *Transportation Science* 27, 211–227.

Vickrey, W. (1969). Congestion theory and transport investment. *American Economic Review* 59, 251–260.

Vossen, T., Ball, M.O. (2006a). Optimization and mediated bartering models for ground delay programs. *Naval Research Logistics* 53, 75–90.

Vossen, T., Ball, M.O. (2006b). Slot trading opportunities in collaborative ground delay programs. *Transportation Science* 40, 29–43.

Vossen, T., Ball, M.O., Hoffman, R., Wambsganss, M. (2003). A general approach to equity in traffic flow management and its application to mitigating exemption bias in ground delay program. *Air Traffic Control Quarterly* 11, 277–292. A preprint was published in *Proceedings of 5th USA/Europe Air Traffic Management R&D Seminar*, 2003.

Vranas, P.B., Bertsimas, D.J., Odoni, A.R. (1994). The multi-airport ground holding problem in air traffic control. *Operations Research* 42, 249–261.

Wambsganss, M.C. (1996). Collaborative decision making through dynamic information transfer. *Air Traffic Control Quarterly* 4, 107–123.

Wei, G., Yu, G., Song, M. (1997). Optimization model and algorithm for crew management during irregular operations. *Journal of Combinatorial Optimization* 1, 80–97.

Willemain, T.R. (2002). Contingencies and cancellations in ground delay programs. *Air Traffic Control Quarterly* 10, 43–64.

Wilson, F.W. (2004). A stochastic air traffic management model that incorporates probabilistic forecasts. In: *Proceedings of 20th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*.

Wyndemere, I. (1996). An evaluation of air traffic control complexity. Report NASA-2-14284, NASA Ames Research Center, Moffett Field, CA.

Yan, S.Y., Lin, C.-G. (1997). Airline scheduling for the temporary closure of airports. *Transportation Science* 31, 72–82.

Yan, S.Y., Tu, Y.P. (1997). Multifleet routing and multistop flight scheduling for schedule perturbation. *European Journal of Operational Research* 103, 155–169.

Yan, S.Y., Yang, D.H. (1996). A decision support framework for handling schedule perturbation. *Transportation Research B* 30, 405–419.

Yan, S.Y., Young, H.F. (1996). A decision support framework for multi-fleet routing and multi-stop flight scheduling. *Transportation Research A* 30, 379–398.

Yen, J., Birge, J. (2000). A stochastic programming approach to the airline crew scheduling problem. Technical report, University of Washington, Seattle, WA.

Young, H.P. (1994). *Equity in Theory and Practice*. Princeton Univ. Press, Princeton, NJ.

Yu, G. (1995). An optimization model for airlines' irregular operations control. In: *Proceedings of the International Symposium on Operations Research with Applications in Engineering, Technology, and Management*, Beijing, China.

Yu, G. (Ed.) (1997). *Operations Research in the Airline Industry*. Kluwer Academic, Boston, MA.

Yu, G., Arguello, M., Song, G., McCowan, S., White, A. (2003). A new era for crew recovery at Continental Airlines. *Interfaces* 33, 5–22.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14002-5

Chapter 2
Public Transit

Guy Desaulniers

Department of Applied Mathematics and Industrial Engineering, École Polytechnique and GERAD, Montréal, Québec, H3T 2A7 Canada
E-mail: Guy.Desaulniers@gerad.ca

Mark D. Hickman

Department of Civil Engineering and Engineering Mechanics, The University of Arizona, Tucson, AZ 85721, USA
E-mail: mhickman@engr.arizona.edu

1 Introduction

For several decades now, operations research has been successful for solving a wide variety of optimization problems in public transit. Several commercial software systems based on operations research techniques have been designed and used by the transit agencies to help them plan and run their operations. Operations researchers have been attracted by the public transit problems because of their size and complexity. Indeed, some of them are huge in practice. For instance, the New York City Transit Authority employs more than 12,000 drivers to operate approximately 4500 buses that serve over 240 bus routes. Furthermore, these problems are complex because they involve passengers, buses, and drivers that are subject to individual preferences and constraints, and interact with each other according to a set of prescribed relationships.

The main goal of most transit agencies is to offer to the population a service of good quality that allows passengers to travel easily at a low fare. The agencies thus have a social mission which aims at reducing pollution and traffic congestion, as well as increasing the mobility of the population. In most cases, the goal is usually not to make profits, as is the case for almost all other transportation organizations such as airlines, railroads, and trucking companies. They are, however, subject to budgetary restrictions that force them to manage expensive resources such as buses, drivers, maintenance facilities, and bus depots as efficiently as possible. Briefly stated, the global problem faced by the agencies consists of determining how to offer a good-quality service to the passengers while maintaining reasonable asset and operating costs.

Addressed as a whole, this global problem is not tractable. Hence, it is divided into a set of subproblems that are usually solved sequentially at various stages of the planning process (strategic, tactical, and operational), and even during operations (real-time control). Strategic planning problems con-

cern long-term decisions such as the design of the transit routes and networks. Most of these problems fall within the category of network design problems and require solving passenger assignment problems as subproblems or for evaluation purposes. These strategic problems aim at maximizing service quality under budgetary restrictions. Tactical planning problems concern the decisions related to the service offered to the public, namely the frequencies of service along the routes and the timetables. These problems are usually solved on a seasonal basis, with occasional updates. These problems also focus on the quality of service. Operational planning problems relate to how the operations should be conducted to offer the proposed service at minimum cost. They include a wide variety of problems such as vehicle scheduling, driver scheduling, bus parking and dispatching in garages, and maintenance scheduling. These problems are solved at various intervals that range from once per month for driver scheduling to once per day for bus parking and dispatching. In contrast to the objectives of the previous problems, the objective for the operational planning problems is clearly one of minimizing total cost. Finally, real-time control problems manage perturbations to the plan using several control strategies. These problems are solved in real time during operations and aim at minimizing passenger inconvenience. Usually, minimal or no operating costs are involved since they consider minor perturbations to the scheduled service.

The goal of this chapter is to review state-of-the-art models and approaches for solving these public transit problems. This review is not exhaustive as it mostly covers the recent contributions that have been applied or have the potential to be applied from our viewpoint. Readers interested on earlier works are referred to the survey paper by Odoni et al. (1994), as well as the series of books arising from the Computer-Aided Transit Scheduling conferences that have contributed tremendously to the practice and growth of operations research in public transit. These books are listed at the beginning of the references, i.e. Wren (1981), Rousseau (1985), Daduna and Wren (1988), Desrochers and Rousseau (1992), Daduna et al. (1995), Wilson (1999), Voss and Daduna (2001), and Hickman et al. (in press).

In certain cases, computational results are briefly reported to give an idea of the problem sizes that can be solved. However, these results are not intended to compare the different approaches. In fact, they can hardly be compared because they most of the times have been obtained using different computers or on different datasets that did not necessarily exhibit the same characteristics.

2 Strategic planning

At the strategic level, transit planning is concerned with the design of transit routes and networks. This involves designing a network of routes to meet passenger demand. Since the demand is based in large part on the network design, the network design problem relies heavily on methods to determine

passengers' route choices (or "assignment" to routes) serving their origins and destinations. This section includes a description of the transit network design problem and a discussion of research in passenger assignment.

2.1 Network design

The public transit network design problem is somewhat more complicated than the traditional network design problem. In addition to determining what links to include in the network, the transit network design includes assembling these links into fixed routes, and determining the frequency of service on each route. The result of the network design, then, should include a set of routes and their frequencies. Most commonly, the problem is formulated on a graph with nodes, links, and (subsequently) routes. Let $G = (N, A)$ be a graph with N, the set of nodes, and A, the set of links, and R represents the set of routes. Nodes represent intersections (e.g., road intersections), but can also represent zone centroids where a geographic zone is represented by a single point (the centroid). A link between nodes represents a particular mode of transport between nodes, and a route represents a sequence of nodes and links of a single mode.

As input to this problem, the formulation typically assumes that there is an existing origin–destination (O–D) matrix, covering the demand between a set of nodes or zones, either on a daily basis or for a specific period within the day. Alternately, it is reasonable (although complicated) to assume that demand is endogenous and determined as an equilibrium problem, in which the flows are a function of the network design. With the origin–destination flows and an assignment of these flows to routes, the set of routes and their frequencies must be determined.

One of the challenges of network design is in the specification of the objective function. Most commonly, the objective is to minimizing the total travel time or the generalized cost of travel. The generalized cost may be found by applying different weights in the objective function to the different components of travel time such as walking (or access) time, initial waiting time, in-vehicle time, transfer time, and egress time separately. Some formulations also include the number of transfers as a component in the generalized cost.

In addition, the costs to the transit operator may also be considered, either explicitly in the objective function or through a constraint on the total budget (or operating profit or loss). Such costs can include the operating cost, given as a function of the route length (in distance or time) and frequency, and the fixed cost of a bus fleet and/or infrastructure along the route network (for rail transit networks). If operator costs are included, a composite objective may be formulated, or the problem can be specified as a multiobjective programming problem.

In addition to this traditional formulation of the objective, several other constraints often enter into the problem in practice; these include: (1) ensuring adequate coverage in the network to provide access to specific nodes or zones

in the service area; (2) ensuring minimum frequencies of service to specific nodes or links in the network; and (3) any other design considerations such as the availability of infrastructure or right-of-way for routes.

In practice, the network design problem incorporates each of these objectives and constraints through a more interactive formulation of the problem, where sample routes may be constructed by computer methods, but are ultimately selected in conjunction with manual review by network designers and planners. In this review, we discuss some of the more significant methods that rely heavily on mathematical programming techniques.

The network design problem is known to be NP-hard (Magnanti and Wong, 1984). As a result, approaches to the problem rely on heuristic techniques to solve problems of reasonable size. Much of the initial work decomposed the problem into two stages: in the first, the set of routes is constructed; in the second, the set of frequencies for these routes are determined. This includes the work of Lampkin and Saalmans (1967) and Silman et al. (1974). In the first stage, heuristic methods are used to construct "skeleton" routes, and these skeleton routes are expanded to cover the full set of nodes in the network. Once the routes are defined, the frequencies are determined by minimizing the total passenger travel time, calculated as the sum of the O–D demand D_{ij} multiplied by the travel time T_{ij}, subject to a constraint on the total fleet size (as a budget constraint):

$$\text{minimize} \quad \sum_{i \in N} \sum_{j \in N} D_{ij} T_{ij}(\mathbf{f}) \tag{1}$$

subject to:

$$\sum_{r \in R} \lceil RT_r f_r \rceil \leqslant \text{total fleet size}. \tag{2}$$

In this formulation, RT_r is the round-trip time on route r and f_r is the frequency on route r. The travel time T_{ij} includes the expected waiting time, as a function of frequencies $\mathbf{f}$ of routes serving the origin i and any transfer node k on the shortest path serving the O–D pair i, j. Lampkin and Saalmans (1967) used a random gradient-based search procedure to determine the final frequency values. In Silman et al. (1974), a penalty is added to the objective for the estimated number of standees on the bus; this penalty is given as a piecewise differentiable function of the route frequency. Their approach uses a gradient projection method to minimize the total travel time.

Dubois et al. (1979) decomposed the network design problem into three subproblems. The first involves determining the links in the street network on which to operate the service; the second determines the routes themselves; and the third determines the optimal frequencies on each route. In the first step, a traditional network design problem is formulated, where the objective is given in (1), using only in-vehicle times as the travel times, subject to a budget constraint on the cost of operating on a street, and binary decision variables indicating whether a street segment is in the final solution. A heuristic is used

to solve this problem, beginning with an initial spanning tree to minimize the total travel time and adding links to minimize this total. A simple all-or-nothing assignment of the O–D flow on the shortest paths is used to estimate the objective function. In the second step, a maximal set of routes is generated from the street network. This set of routes is then subject to heuristic rules to determine the final route structure: (1) routes with heavy transfer flow are joined; (2) route segments are deleted where the demand is effectively served by other routes; and (3) routes that overlap are joined. In the third step, the optimal route frequencies are found using a gradient-based search heuristic, similar to Lampkin and Saalmans (1967). Waiting times are explicitly considered in this last step.

A more formal presentation of the transit network design problem, from a mathematical programming approach, is given by Hasselström (1981). Hasselström proposed a two-stage process of network design in which routes and frequencies are determined simultaneously. In the first stage, an initial route network is generated; in the second stage, this route network is refined and a detailed evaluation of the routes and the passenger assignment is performed. In addition to solving the route and frequency problem simultaneously, the advantage of Hasselström's method is in its implementation and application for realistic network sizes.

In the first stage, Hasselström's (1981) formulation includes a direct demand function, allowing the demand to be determined endogenously. The form of the direct demand model is based on a traditional gravity model with parameter β, where all terms not dependent on the route structure or the frequencies are rolled into a constant term K_{ij} for each O–D pair i, j. Remaining elements of the generalized cost are given by C_{ij}, which is a function of the set of frequencies $\mathbf{f}$. The frequency of each route r is denoted f_r. The objective function, maximizing consumer surplus, is equivalent to maximizing the number of passengers with this demand function. As constraints, Hasselström (1981) considered a budget constraint $\overline{C}$ that includes a cost per vehicle on each route c_r. Also, there is a required minimum frequency of service for a zone $s \in S$, given as Δ_s. If a route r serves zone s, this is indicated with a binary parameter δ_{rs}. Finally, the frequencies must be included in a feasible set $\mathcal{F}$ (e.g., nonnegative integers).

The network design problem in the first stage is then formulated as follows (in this formulation, the notation is simplified to deal with a single transit mode; multiple modes may also be considered in the network):

$$\text{maximize} \quad \sum_{i \in N} \sum_{j \in N} D_{ij} \tag{3}$$

subject to:

$$D_{ij} = K_{ij} \mathrm{e}^{-\beta C_{ij}(\mathbf{f})}, \quad \forall i, j \in N, \tag{4}$$

$$\sum_{r \in R} c_r f_r \leqslant \overline{C}, \tag{5}$$

$$\sum_{r \in R} f_r \delta_{rs} \geqslant \Delta_s, \quad \forall s \in S, \tag{6}$$

$$f_r \in \mathcal{F}, \quad \forall r \in R. \tag{7}$$

In (4), the generalized cost term C_{ij}, a function of the frequencies $\mathbf{f}$, includes waiting and transfer times for the O–D pair, as determined in the passenger assignment. The demand is therefore given as a function of the service frequencies.

To solve this model in the first stage, an initial network is generated by enumerating all possible routes serving a pair of terminals. A set of heuristic rules is then used to prune clearly inferior routes from this set. Then, the final routes and frequencies are constructed by solving a mathematical program: one method uses a linear program to maximize the passenger flow; a second method uses a convex nonlinear program to maximize the consumer surplus. In this first stage, the assignment uses all common routes (see Section 2.2.1) to determine waiting and transfer times. The decision variables in both formulations are the frequencies of each route; routes with very low frequencies can be pruned from the solution space.

In the second stage, a detailed assignment is performed, and the routes are refined. Passenger assignment on existing routes is performed with the heuristic of Andreasson (1977) (discussed in Section 2.2.1). With this new assignment, several route refinements are considered. First, optimization of the connection of route segments at route intersections is proposed; this problem is formulated as a maximum weighted matching problem of combining route segments at the point of intersection. Also, a nonlinear program is formulated for re-optimization of frequencies, using the vehicle fleet size constraint. This optimization is solved by Lagrangian relaxation.

Hasselström (1981) reports on a case study with 50 local bus routes, 10 tram routes, and express bus service. Since this time, the methodology has been developed as commercial software, and hence is clearly able to solve realistic problem sizes.

In the work of Ceder and Wilson (1986), two different mathematical formulations of the bus network design problem are suggested. The first formulation considers the passenger objective of minimizing excess travel time upon boarding, expressed as the sum of "excess" travel time (larger than the shortest travel time with a direct route) plus the transfer time (if any), summed across all O–D pairs. This objective is minimized subject to constraints on the maximum O–D travel time (as a percentage above the shortest path), lower and upper bounds on the route length (expressed in units of running time), and a constraint on the maximum number of routes. A second formulation adds the passenger waiting time and vehicle operating and capital costs to the objective function; it also includes constraints on the minimum frequency for each route and a constraint on the maximum fleet size. To generate a large set of feasible routes, Ceder and Wilson (1986) proposed a heuristic in which each designated terminal node is processed separately. A (topological) breadth-first

search is conducted, in which potential routes that do not meet the constraints on the maximum travel time are eliminated. In addition, the total "excess" passenger hours are also calculated for the route. This set of feasible routes can then be used for further screening and passenger assignment.

The mathematical formulation of Ceder and Wilson (1986) was extended more recently by Israeli (1992) and related papers (Israeli and Ceder, 1989, 1995; Ceder and Israeli, 1998). In this research, the transit network design problem is formulated as a multiobjective programming problem, with two objectives: the total passenger cost (Z_1) and the operator fleet size (Z_2). The total passenger costs Z_1 includes the in-vehicle passenger hours spent between the origin and destination PH_{ij}, the waiting and transfer time spent traveling from the origin to the destination WH_{ij}, and the empty seat-hours on a route r denoted EH_r. These three terms are weighted (weights a_1, a_2, and a_3) in the objective. Formally, the objectives are described as

$$\text{minimize} \quad Z_1 = a_1 \sum_{i \in N} \sum_{j \in N} PH_{ij} + a_2 \sum_{i \in N} \sum_{j \in N} WH_{ij} + a_3 \sum_{r \in R} EH_r, \quad (8)$$

$$\text{minimize} \quad Z_2 = \text{fleet size}. \quad (9)$$

Constraints in the formulation include the passenger assignment from a fixed demand matrix, and minimum frequencies on each route. This problem is solved with the following heuristic:

1. The full set of feasible routes is enumerated, in a manner similar to Ceder and Wilson (1986).
2. Additional direct routes are added to the network between O–D pairs with high demand, where the origin and destination nodes are not terminals. Second, the number of transfers required for each O–D pair for the given route structure is calculated, up to a maximum (e.g., 1 or 2 transfers).
3. A minimal set of routes is obtained through a heuristic procedure. This problem is set up as a set covering problem with the full set of feasible routes. Each column is defined as a route or a feasible combination of routes meeting the maximum number of transfers; the rows are O–D pairs. The objective minimizes the deviation from shortest paths, while maintaining constraints on connectivity for each O–D pair (i.e., reachable within the maximum number of allowable transfers).
4. The assignment of flow to paths and the frequencies on each path are determined iteratively. The frequencies are determined based on the peak load segment on each route, and these in turn are used in calculating the waiting and transfer times in the assignment. The assignment procedure is loosely based on that of Marguier and Ceder (1984), described in Section 2.2.1. With this information, Z_1 is calculated.
5. The minimum fleet size Z_2 is determined using the method of Stern and Ceder (1983).

6. New routes are considered to explore other points in the solution space for the two objectives. In this procedure, a column generation technique is used to avoid re-evaluating previously accepted route sets. These new route sets are re-evaluated by repeating steps 3–5.
7. The route sets in the efficient frontier are evaluated and presented to the decision-maker.

The example problem and solution in Israeli and Ceder (1995) is a problem with 8 nodes and 14 links (based on a similar problem from Ceder and Wilson (1986)). It is unknown how the solution method would perform on larger, more realistic networks.

Another mathematical programming approach to the transit network design problem was proposed by van Nes et al. (1988). In this approach, the routes and frequencies are determined simultaneously. The objective function maximizes the number of direct trips (i.e., trips without transfers) served in the network, for a given fleet size. A direct demand model is proposed to estimate the origin–destination trips by public transit; trips are proportional to the attraction of the origin zone i, O_i, and the destination zone j, D_j, and is an exponential function of the cost, similar to that of Hasselström (1981). The objective function is formulated as

$$\text{maximize} \quad \sum_{i\in N}\sum_{j\in N} aO_iD_j\mathrm{e}^{-\beta C_{ij}(\mathbf{f})}. \tag{10}$$

The generalized cost term C_{ij} is defined as an explicit function of the frequencies of the optimal subset of routes serving the passenger's origin i, R_i^* (Chriqui and Robillard, 1975):

$$C_{ij}(\mathbf{f}) = K_{ij} + \frac{60\alpha}{\sum_{r\in R_i^*} f_r} + c. \tag{11}$$

In this equation, α and c are constants. This equation assumes, for direct service, a constant K_{ij} for access time, egress time, and in-vehicle travel time, and adds a term for the waiting time as a function of the frequencies on acceptable routes.

A variety of constraints are used in this formulation. For these, consider a set $M = \{1, \ldots, m\}$ of vehicle types, with N_m the total number of vehicles available of type m. Operating a vehicle of type m incurs a cost factor k_m. A total of N_r vehicles are assigned to route r, with the indicator b_{mr} equal to 1 if vehicle type m is assigned to route r. With these variables, the constraints include: a budget constraint of $\overline{C}$, the vehicle availability N_m, the set of feasible frequencies $\mathcal{F}$, and the allocation of buses among routes based on the frequency and the round-trip time on the route RT_r. Mathematically, these constraints are formulated as:

$$\sum_{m\in M} k_m \sum_{r\in R} N_r b_{mr} \leqslant \overline{C}, \tag{12}$$

$$\sum_{r \in R} N_r b_r \leqslant N_m, \quad \forall m \in M, \tag{13}$$

$$f_r \in \mathcal{F}, \quad \forall r \in R, \tag{14}$$

$$N_r - 1 < f_r \frac{RT_r}{60} \leqslant N_r, \quad \forall r \in R. \tag{15}$$

The solution technique adopted by van Nes et al. (1988) is a heuristic in which all frequencies on proposed routes are set to 0. Each route is evaluated with respect to its potential to improve "efficiency", defined as the ratio of passengers added by the direct service to the additional cost of increasing the frequency, evaluated in (12)–(15). The route with the highest efficiency is selected and the frequency on that route is increased, until the budget and the available vehicles are consumed. It is shown that this heuristic is similar to evaluating the Kuhn–Tucker conditions for the problem when the budget constraint is included in the objective with a Lagrange multiplier. The Lagrange multiplier is identical to this "efficiency" measure, and these should be approximately equal across routes in the final solution.

van Nes et al. (1988) report testing this solution technique on a network from the Netherlands with 182 nodes and 115 zones, for a network of 8 routes. The paper also reports that the modeling system is capable of solving instances up to 250 nodes, 150 zones, and 750 possible routes.

More recent work has also included a number of metaheuristic methods. Baaj and Mahmassani (1995) and related work (Baaj and Mahmassani, 1990, 1992) decompose the network design problem into three elements: a route generation step, in which routes and frequencies are constructed; a network analysis procedure, defining measures of effectiveness at the network-, route-, and stop-level; and a route improvement algorithm to improve the route design. The heuristic proposed by Baaj and Mahmassani (1995) begins by generating additional skeleton routes connecting the highest O–D pairs in the demand matrix with direct service. With these skeletons, a set of possible node selection and insertions strategies are used to generate full routes. Then, passenger assignment follows the method of Han and Wilson (1982) (see Section 3.1) and determines the frequency and number of buses on each route according to a pre-specified maximum loading factor. Once this assignment is completed, the network evaluation tool is used to identify the number of trips satisfied by direct, one-transfer, and two-transfer trips, and the total waiting time, in-vehicle travel time, and transfer time in the network. The route improvement procedures then are called to improve the route structure through heuristics that: (1) prune off low ridership routes and/or route segments and joining these with other routes; and (2) consider improvements to routes by splitting routes into two parts or by exchanging route legs between routes at points of intersection.

A recent study by Fan and Machemehl (2004) examined simulated annealing, tabu search, genetic algorithms, local search, and random search techniques to solve the network design problem. As with other previous methods,

all the techniques begin with a set of skeleton routes. These metaheuristics are used to generate additional routes; the output is run through a network evaluation tool. Subsequent iterations between the network analysis tool and the metaheuristics are used to improve the quality of the solution.

Other authors have investigated the use of genetic algorithms for the transit network design problem; these include (among many others) recent works by Pattnaik et al. (1998), Bielli et al. (2002), Tom and Mohan (2003), and Verma and Dinghra (2005). These methods involve two steps. First, all feasible routes are generated, typically in a method similar to that proposed by Ceder and Wilson (1986). A set of routes are coded into the genetic algorithm as a string, containing a certain number of routes. For this technique, a fixed number of routes are required by the algorithm, although the number of potential routes can vary so as to determine the optimal number of routes. These strings are then evaluated using the assignment and network evaluation techniques of Baaj and Mahmassani (1995), and consistent with the methods of genetic algorithms, the pool of strings to be evaluated is evolved to a new population, and the process iterates. The size of networks in the genetic algorithm approach can be somewhat larger than with the analytic methods; Bielli et al. (2002) report solving an instance of 1134 nodes, 3016 arcs, 459 stops, and 22 routes. Tom and Mohan (2003) report solving an instance with 1332 nodes (bus stops) and 4076 arcs.

2.2 Passenger assignment

One of the critical issues in strategic planning is determining the demand on each route and other measures of service consumption. Most of the more common strategic measures of performance, from the perspective of the passenger, relate to the amount of time and money spent traveling in the network; i.e., elements of the passenger's path from the origin to destination. Hence, the passenger assignment is critical to determining system performance.

The passenger assignment problem can be defined as follows. Given an origin-to-destination flow, what are the flows on paths through the transit network, taken by the passengers? In formulating this problem, the passenger's objective is assumed to be minimizing travel time or generalized cost. The travel time may consist of some or all of the following variables, with perhaps different weights: the time to access a stop, the waiting time in the stop, the in-vehicle time, the transfer time, the number of transfers, egress time, and any monetary cost. The passenger then faces the task of selecting a route or set of routes that may be able to get from the origin to the destination with the minimum time or generalized cost. In the literature, this problem is addressed within the network design problem (Section 2.1), as part of the task of determining frequencies on routes (Section 3.1), and also as a unique problem itself.

In contrast to the simplicity of the problem definition, there are a number of aspects of the problem that have led to several different research approaches.

One important concept in passenger assignment is the determination of the "minimum cost" path. Important elements in defining the attributes of cost include:

1. The characterization of time-dependence and stochastic attributes in the minimum cost path.
2. The characterization of a solution as: (1) a *single path*, including only a route or combination of routes; (2) a path that can include a set of *common lines*, including cases where multiple routes may overlap on some part of the shortest path; or (3) a *strategy*, allowing passengers to choose their own boarding rules as they travel from origin to destination.
3. The effect of capacity and crowding in the transit network.

In the characterization of the minimum cost path, a traditional approach assumes that deterministic values can be used for travel times on each link. Traditional shortest path techniques have been easily modified to solve these problems. More recent approaches have included stochastic and time-dependent features of the travel time: passenger arrivals, vehicle arrivals, and travel times may be stochastic. As might be expected, these stochastic processes will greatly affect the path assignment approach (Nuzzulo, 2003). There is considerable evidence that passenger arrivals appear to be Poisson for higher-frequency (lower-headway) service, with headways up to 10–15 minutes. In these cases, the assumption of Poisson passenger arrivals appears to be common. However, at longer headways (lower frequencies), some fraction of passengers may actually time their arrivals with the schedule, which may again significantly complicate the analysis (Turnquist, 1978; Bowman and Turnquist, 1981). Moreover, the treatment of vehicle arrivals may be considered deterministic (according to schedule or at the given headway) or stochastic. If vehicle arrivals are assumed to be Poisson, many of the calculations in the path assignment simplify considerably.

The element of time-dependence, relating to the transit schedules, can also affect the modeling approach. In its simplest case, with perfect adherence to the schedule, the choice of a "minimum cost" path decomposes into a time-dependent shortest path problem. This is usually well solved using variants of existing shortest path techniques; see, for example, Tong and Richardson (1984). However, when some combination of both time-dependence and stochastic travel times are introduced, the problem is not so well behaved. As was shown by Hall (1986), the problem of finding a stochastic and time-dependent shortest path suffers from the fact that subpaths do not necessarily concatenate; instead, a possibly exponential number of paths must be evaluated to ensure an optimal path is found. When vehicle arrivals are random but somewhat correlated with a schedule, the assignment becomes significantly more complicated (Hickman and Bernstein, 1997).

A second complication for transit networks is that there may not be a single route or set of routes which has the minimum cost. This may occur in cases where multiple transit routes may overlap on some part of the origin–

destination path. This problem is commonly referred to as the *common lines* problem, in which a passenger may take one of many routes for at least part of the path from the origin to the destination. The more general case of multiple origin–destination paths has led to the term *strategies* (Spiess and Florian, 1989), reflecting possible boarding rules the passenger may use in traveling from an origin to a destination. In a graph-theoretic model, the subnetwork of eligible paths from the origin to the destination in a strategy is characterized as a *hyperpath* (Nguyen and Pallottino, 1988).

The final characterization that may be made is based on the treatment of capacity and crowding (de Cea and Fernández, 1996, 2000). Much of the early literature in the passenger assignment assumed that vehicle capacity was not typically exceeded, and as a result, capacity and crowding effects could safely be ignored. This allowed certain simplifications of the problem, although this is clearly not applicable in all circumstances. Rather, if passenger volumes are assumed to run close to or over the capacity of a route, it might be expected that passengers may not be able to board the first vehicle to arrive. Hence, waiting time and transfer times may be directly affected by the volume of passengers on the route, creating congestion effects. This congestion affects the problem formulation and solution techniques. As a result, the discussion that follows in Sections 2.2.1 and 2.2.2 is decomposed into the passenger assignment under "uncongested" conditions and "congested" conditions, respectively.

2.2.1 Uncongested assignment

The earliest methods of transit assignment used variants of well-known shortest path algorithms; examples include Dial (1967), Lampkin and Saalmans (1967), le Clercq (1972), Silman et al. (1974), and Last and Leak (1976). In these cases, the full demand of each O–D pair is assigned to a shortest path. The variations from existing shortest path methods are based on two exceptions: waiting times and common lines in the network. These two issues are intertwined. A general formulation of the waiting time, as a function of the frequency, suggests that waiting time is related to the inverse of the frequency of routes serving the passenger. If R_i is the set of routes serving the stop i that also serve the passenger's destination or an intermediate (transfer) node, then the expected waiting time is given as

$$E[WT] = \frac{\alpha}{\sum_{r \in R_i} f_r}, \tag{16}$$

where α is a parameter, such that $\alpha = 1$ for Poisson vehicle arrivals, $\alpha \approx 0.5$ with deterministic arrivals. With the expression in (16), the shortest path problem can be solved by creating additional links in the network representing the corresponding waiting time. Moreover, if more than one route serves a node i and also serves a given intermediate node or destination node for an O–D pair, the assignment is made to each route on the basis of the frequency share. That

is, the fraction of passengers served by route r', $P_{r'}$, is given as

$$P_{r'} = \frac{f_{r'}}{\sum_{r \in R_i} f_r}. \tag{17}$$

This formulation assumes that the passenger takes the first bus to arrive at a stop, among all routes serving that stop.

Chriqui and Robillard (1975) provided a more rigorous treatment of the problem when "common lines" serve some part of an O–D path. Specifically, it may not be to the passenger's advantage to choose the first among all routes serving the pair of stops. The most notable case is where one or more of the routes has a shorter travel time to the destination than the others. In this case, it may be advantageous not to board a bus on a slow route, if it is the first to arrive. Chriqui and Robillard (1975) formulated this problem as follows. Assume the passenger will choose a subset of routes, and will board the first route of this subset to arrive at the origin stop. One may also assume that the travel time to the destination after boarding is constant for each route, but may vary across the set of routes serving the stop. Finally, we assume that the passenger desires to minimize the sum of waiting time and time after boarding. Then, the selection of routes becomes a hyperbolic programming problem of selecting routes to include in this subset. Practically, this problem can be solved to optimality by enumerating the possible route subsets. The result is an optimal route subset R_i^* that can be used to define the waiting time at the node and frequency shares for each route in the subset, using the subset R_i^* in the summations of (16) and (17). Chriqui and Robillard (1975) also derived expressions for waiting times and frequency shares for both uniform- and exponentially-distributed bus arrival times. These results were extended by Marguier and Ceder (1984) and Israeli and Ceder (1996), in which routes can be grouped into slow routes and fast routes.

A related heuristic approach was suggested by Andreasson (1977), in which a route is considered in the desirable subset if the travel time upon boarding a bus on that route is less than or equal to the waiting time plus the travel time upon boarding of the minimum time route. Waiting times and route shares are then determined based on this route set. Andreasson's method was also extended by Jansson and Ridderstolpe (1992), in which they present an iterative heuristic for calculating waiting times and route proportions for transit networks with multiple routes and modes between an O–D pair. Jansson and Ridderstolpe (1992) showed that the functions (16) and (17) can create poor approximations of the waiting time and route shares under deterministic headways; instead, these functions depend heavily on the exact timetable.

The work of Spiess (1983) and Spiess and Florian (1989) introduced the concept of passenger *strategies*. In their formulation, the passenger is assumed to minimize the sum of waiting time and on-board time, where these may vary based on the passenger boarding rules. They formulate this assignment as one to minimize the total travel time in the network, taken as the product of arc

flows and their associated travel times, plus the total waiting time in the network. This waiting time depends on the routes in each strategy. For a given node i, let A_i^+ be the set of arcs leaving i and A_i^- be the set of arcs entering i. Let v_a be the flow on arc a, t_a be the travel time on arc a, and f_a is the frequency of service on arc a. Also, ω_i is the total waiting time experienced by passengers boarding at node i. The demand generated at node i is g_i, and V_i (a parameter) is the total flow entering node i. The formulation of the passenger assignment problem is given as an integer linear program, in which the decision variables x_a are binary variables indicating if an arc a is in the strategy. The relaxation of this problem is

$$\text{minimize} \quad \sum_{a \in A} t_a v_a + \sum_{i \in N} \omega_i \tag{18}$$

subject to:

$$\sum_{a \in A_i^+} v_a - \sum_{a \in A_i^-} v_a = g_i, \quad \forall i \in N, \tag{19}$$

$$\omega_i = \frac{V_i}{\sum_{a \in A_i^+} f_a x_a}, \quad \forall i \in N, \tag{20}$$

$$v_a \leqslant f_a w_i, \quad \forall a \in A_i^+, \forall i \in N, \tag{21}$$

$$v_a \geqslant 0, \quad \forall a \in A, \tag{22}$$

$$0 \leqslant x_a \leqslant 1, \quad \forall a \in A. \tag{23}$$

The dual of this linear program has the form of a shortest path problem, resulting in an optimal label-setting algorithm for its solution.

A similar formulation of the assignment problem is presented by de Cea and Fernández (1989), with the use of nonlinear equality constraints for (21). The resulting nonlinear optimization problem is solved by incorporating the nonlinear constraints into the objective function. The solution technique can then be decomposed into three parts: the selection of common lines (a hyperbolic programming problem) for each O–D pair; the assignment of the O–D volumes to links representing common routes in the hyperpath; and the assignment of common route flows to specific routes by frequency share.

More recent study of passenger assignment has focused on the assignment of passengers to specific scheduled vehicle trips. That is, the assignment identifies a particular vehicle trip to which a passenger is assigned. In this formulation, it is necessary to have a time-dependent origin–destination matrix $D_{ij}(t)$. An extensive discussion of various approaches to schedule-based transit assignment, both for uncongested and congested transit networks, is found in the recent volume edited by Wilson and Nuzzolo (2004).

In Tong and Wong (1999), the time-dependent shortest path technique of Tong and Richardson (1984) is used to assign trips to the network. Additional complications are added using stochastic weights on the various components

of travel time (walk time, waiting time, and transfer time), relative to in-vehicle time.

A combination of stochastic and time-dependent travel time attributes are included in Hickman and Bernstein (1997). This model characterizes the following passenger behavior: upon arriving at a stop, the passenger waits until a bus arrives, and then determines whether or not to board the bus based on the time spent waiting and the set of additional bus arrivals expected in the future. The formulation of this problem essentially requires full enumeration of all paths from the origin to the destination, accompanied by the derivation of the probability distributions of travel times on all paths, at the time the boarding decision is made. Such "clever" passenger behavior can be used to simulate passenger behavior under real-time information, as illustrated by Hickman and Wilson (1995). A similar framework for passenger boarding strategies with information at the stop was also presented by Gentile et al. (2005).

2.2.2 Congested assignment

The area of congested transit assignment has evolved in examining the effects of capacity limitations on passenger path assignment. With vehicle capacities, it may be the case that demand is sufficiently high that a passenger desiring to board a given vehicle may be unable to. This results in higher waiting times when in such crowded conditions. This has led to formal specification of *equilibrium* transit assignment in which the delay that each passenger imposes on other passengers is explicitly included in the model. In general, the effect of crowding and vehicle capacity is incorporated in the modeling through the impact of additional waiting and/or transfer time caused by passengers being unable to board a desired vehicle because it is full. It may also be included as additional "discomfort" experienced by passengers while on board the vehicle. However, in these models it is important to note that the effect of congestion is clearly not symmetrical by arc flows.

Mathematically, the most common approach to include capacity and crowding has been to formulate the waiting time as an increasing function of the volume on a particular line, both in terms of the passengers on-board and the passengers waiting to board at a stop. A graph-theoretic structure for transit equilibrium assignment was developed by Nguyen and Pallottino (1988). These authors proposed the graph concept of a *hyperpath*, defined as an acyclic, directed subgraph of routes connecting the passenger's origin–destination pair. Such a hyperpath may be defined using a particular passenger *strategy* (Spiess and Florian, 1989). Passengers are assumed to travel on the *shortest hyperpaths* connecting their origin and destination.

When equilibrium conditions occur, the problem formulation and solution methods from traditional traffic assignment can be used (refer to Chapter 10 in this volume). Nguyen and Pallottino (1988) suggested using traditional traffic assignment techniques, with the adaptation of these techniques to calculate shortest hyperpaths (or strategies) rather than shortest paths, when calculating a direction for improvement in the assignment.

In de Cea and Fernández (1993), the waiting time is considered to be a function of both the passengers desiring to board at the given stop and those passengers traveling through the stop on board the route. Mathematically, the passenger travel time is calculated as a power function of the *conflicting volume* divided by the capacity to approximate the cost of congestion. The conflicting volume is the sum of the boarding volume and the passengers on board. With congestion on the waiting or boarding arcs, a common approach is to define an *effective frequency* f', determined as the average frequency observed by the passenger, assuming he/she may be denied boarding on the first vehicle in his/her strategy. This effective frequency, calculated as the reciprocal of the expected waiting time for a given route, can be sensitive to the level of crowding upon boarding. In turn, the effective frequency f' can be used in the more traditional waiting time and frequency-based assignment approaches found in (16) and the proportional assignment in (17); for a discussion of these issues, see Bouzaïene-Ayari et al. (2001).

de Cea and Fernández (1993) determined the congestion cost using a linear function of the conflicting volume divided by the capacity. The model uses a variational inequality formulation with nonlinear constraints, with assignment to routes based on the effective frequencies. If the assignment to routes is made on the basis of actual route frequencies, rather than the effective frequencies, the problem has linear constraints. In either case, this problem can be solved using diagonalization, making it susceptible to traditional traffic assignment algorithms.

Wu and Florian (1993) and Wu et al. (1994) extended this work to include a formal strategy formulation of the congested assignment problem. The problem is formulated as an asymmetric network equilibrium problem, but with a variable transformation to solve in the space of hyperpath flows rather than route flows. The solution method uses a symmetric linearization (similar to diagonalization), called the linearized Jacobi method, for solving the resulting variational inequality problem.

Moreover, Cominetti and Correa (2001) extended the models of Wu et al. (1994) to consider an *effective frequency* model of transit assignment, considering possible congestion on the boarding links. The principle finding in this work is the definition of some necessary and sufficient conditions that an equilibrium transit assignment has been reached, both in terms of arc flows and in terms of the strategy. A straightforward cost function is established as an objective. The proposed algorithm performs shortest-hyperpath assignment on the inner loop, while using the method of successive averages to update the flows after each iteration.

Lam et al. (1999) presented a stochastic user equilibrium model for passenger assignment. This model assumes a simple bottleneck model for congestion on a link; i.e., the additional delay is linear with the ratio of the conflicting volume to the capacity. A multinomial logit model is used for the selection of paths. The problem is formulated as a nonlinear programming problem with linear constraints. It is shown that conditions on the Lagrange multipliers in the

Kuhn–Tucker conditions can be specified such that the route capacities are not exceeded. Rather, when route capacities are reached, the bottleneck delays are proportional to the Lagrange multipliers. With this observation, the problem is solved using existing solution techniques for the stochastic user equilibrium problem in traffic assignment (see Chapter 10).

The work of Lam et al. (2002) extended this model to a more disaggregate model of route operations. In a more detailed model of stop operations, the total route travel time is affected by the delay in boarding and alighting at the stop. The frequency on the route, in turn, is determined by the number of vehicles divided by the round-trip travel time. In this case, the waiting time is a function of the frequency, but the frequency itself is a function of the delay caused by crowding. As a result, the assignment is then a fixed point in both the space of frequency and boarding and alighting volumes. This fixed-point problem forms an outer iteration on the assignment and is solved using the method of successive averages.

A more general network model has been presented in the work of Lo et al. (2003). This model accommodates nonlinear fare structures and transfers through the use of a *state-augmented multimodal* (SAM) network. In this network representation, the node itself is augmented based on the opportunities to transfer at the node, the number of transfers made by the passenger in the network, and explicit representation of direct links in the network to represent nonlinear costs. This network is then applied in a stochastic user equilibrium assignment, using a logit model. This framework has also been extended to a nested logit structure by Lo et al. (2004).

Nielsen (2000) considered a stochastic user equilibrium model that is based on congestion both for waiting times as well as in-vehicle discomfort. Both measures are functions of the flow on the associated link and the on-board capacity of the vehicle. The stochastic user equilibrium is based on the probit model for the selection of paths, among common lines (line aggregation). Existing methods for probit-based traffic assignment are used to solve the passenger assignment (see Chapter 10). An application using a nested logit model for the stochastic user equilibrium assignment on a large regional transit network is presented in Nielsen (2004).

As with the uncongested assignment, recent work has been examining assignment to specific vehicle trips in the schedule. A transit equilibrium assignment model that explicitly considers schedules was formulated by Nguyen et al. (2001). In their model, the capacity constraints on boarding and transfer links are hard constraints. The additional delay is a function of the *available capacity* of the vehicle as it arrives. The passenger assignment to routes is based on the passengers' desired arrival times at the destination; a penalty term (schedule delay cost) is added to the passenger cost for arriving at the destination at a time other than the desired arrival time. This problem is set up as a nonmonotonic variational inequality problem and solved using simplicial decomposition, in the form of a column generation technique that generates new extreme points in the arc flow solution space.

A separate line of thinking has been developed by Nuzzulo et al. (2001). In these models, the transit passenger is assumed to make a discrete choice of route and trip, based on the attributes of the trip as well as the desired arrival time at the destination. The discrete choice model uses for the utility function typical variables related to travel times, transfers, and related variables. The choice of run is based on trading off these terms in both a within-day assignment and a day-to-day assignment based on learning, which is accomplished through an exponential filter of run attributes. Congestion is included in the model using a measure of delay in the passenger boarding and in-vehicle time.

Also in the area of trip-based assignment, Poon et al. (2004) have extended the work of Tong and Wong (1999) to congested assignment. In this work, a time-dependent origin–destination demand is given, and the assignment is made in a dynamic network based on specific vehicle trips. Congestion is considered through queuing delay at the stop or station as passengers prepare to board, and through the available space on the transit vehicle when it arrives at a stop. The assignment in a single iteration follows Tong and Richardson (1984) using the latest network travel times, and is performed by moving passengers incrementally in time through the network. The final assignment is solved iteratively by the method of successive averages, with the queuing delay and vehicle loading being updated after each iteration.

Finally, a multiagent approach to transit assignment has been developed by Wahba and Shalaby (2005). In this approach, passengers are represented as agents in the transit network. The passenger behavior is described in terms of their route, stop, and departure time choice, based on a desired arrival time at the destination. This behavior is simulated in the network on a given day, and reinforcement learning is used to represent day-to-day adaptation of passengers to their experiences in the network. This is repeated for many days to achieve a final transit network assignment.

3 Tactical planning

In tactical planning, we are concerned with intermediate steps in the planning process in which the frequencies of routes are constructed and the service schedule is determined. We include these two parts together in the tactical level because they are predominantly oriented toward structuring and improving service to the passenger. In this context, Section 3.1 describes the selection of frequencies on the set of transit routes, and Section 3.2 describes methods to construct timetables.

3.1 Frequency setting

The process of determining frequencies for transit routes has already been introduced in the process of network design (Section 2.1). While a set of frequencies are a necessary product of the network design, it is also true that

a transit agency will evaluate and determine frequencies on routes more often than this. Variations in passenger demand patterns and smaller changes in route design may precipitate a need to adjust frequencies. In this section, we begin with the problem of frequency setting for typical routes, and then we briefly describe methods of determining frequencies for some other types of routes that are commonly used.

The problem of setting frequencies can be approached from several different ways. The primary goal is to select frequencies that maximize the passenger service, which can be defined in a number of different ways, subject to a number of possible constraints. These include constraints on the overall fleet size (which is assumed fixed for this process), a constraint that capacity on a route must be sufficient for the demand, and any policy constraints on minimum desirable frequencies. Other input data include the round-trip and required layover time on each route. With this information, a transit agency will choose an allocation of the fleet to particular routes; this allocation will then directly indicate the frequency of service on each route. Finally, these frequencies may be specified by time of day and day of week.

The most common practical approach is to design frequencies to meet the maximum passenger demands without exceeding the capacity, or without exceeding some threshold value of bus utilization, the ratio of demand to capacity (Ceder, 1984). In cases where this produces unreasonably low frequencies, minimum frequencies (or maximum headways) are also commonly applied. More rigorous mathematical programming models have been developed and are outlined below, but these have rarely been applied because of their complexity.

Early work on this problem focused on determining frequencies with common route structures. Scheele (1980) formulated the problem of determining route frequencies as a nonlinear program, based on minimizing the total generalized passenger travel time, with decision variables being the frequency of each route. Simultaneously, this model solves for the flow on each O–D path (the passenger assignment problem). An O–D path is defined strictly as a sequence of route segments. The formulation includes constraints that the demand cannot exceed the available capacity on a route, flow conservation in the assignment, and fleet size constraints. An entropy constraint is also included to distribute trips in the transit network and to ensure accessibility between all origins and destinations. An iterative solution methodology is proposed in which the set of frequencies is fixed, and the O–D and path flows are determined through a Lagrangian function. From the Lagrangian, a descent direction for the frequencies is determined, and the frequencies are updated. The new frequencies are used to iterate on the assignment, until the frequencies converge.

Similarly, Han and Wilson (1982) formulated the problem of solving for frequencies on each route as an allocation of vehicles among routes in a network. The problem is formulated as one of solving for frequencies, with the constraints being the passenger assignment to individual links, the capacity of each

route, and the total fleet size. In contrast to Scheele (1980), however, the objective is to minimize the maximum "occupancy level" at the maximum load point for each route in the network. To solve this problem, Han and Wilson (1982) proposed a two-stage heuristic as follows. In the first stage, a base allocation is achieved that guarantees that all routes have sufficient frequency so that all passengers are served, but there is 100% utilization on at least one route segment for each route. In this first stage, the passenger assignment, O–D pairs are decomposed into those with only a single path (so-called "captive" flow) and those with multipath ("variable" flow) assignment. For the multipath assignment, a simple frequency-share model is adopted for both direct paths and transfer paths. In the base allocation, the captive flow is assigned in the network, and a lower bound on the frequency for each route is determined based on setting the frequency equal to the maximum link flow divided by the vehicle capacity. Second, an iterative procedure is used in which the "variable" flow is assigned and the frequencies are updated to equal the maximum link flow on each route divide by the vehicle capacity. This process iterates until the "variable" flow on each route segment converges. In the second stage, any remaining vehicles in the fleet are allocated to routes to reduce the utilization uniformly. In the example in the paper, this is achieved by increasing the frequency of all routes directly in proportion to the remaining vehicles in the fleet.

Furth and Wilson (1982) presented a model to determine route headways that maximize the consumer surplus (measured in terms of waiting time) plus the total ridership, as a function of the headway. In this formulation, the demand is a function of the headway, making the total ridership and waiting time dependent on the headway. The formulation includes constraints on the total subsidy, the total fleet size, and maximum headway values (as a policy device). The problem is solved through an algorithm using the Kuhn–Tucker conditions on a relaxation where the maximum headway and fleet size constraints are relaxed. Violations of the maximum headway constraint are projected back to the maximum headway, with associated reductions in the available subsidy. Violations of the fleet size constraint are accommodated with a new set of Kuhn–Tucker conditions where this constraint is binding. The algorithm iterates through these conditions until all routes have similar multipliers. The result is an optimal allocation of buses to routes.

A more complex model, minimizing passenger waiting cost, operating cost, and the cost of vehicle "crowding" was introduced by Koutsopoulos et al. (1985). In their formulation, demand is assumed to be fixed, and constraints on the available subsidy, the maximum fleet size, and available capacity on each route. This is formulated as a nonlinear program, which under certain simplifying assumptions is formulated and solved as a linear program.

Recent work by Gao et al. (2004) has explored a bi-level model for determining line frequencies and the corresponding network assignment. The upper level solves for the optimal frequency of each route, minimizing the total passenger cost. This solution is then iterated with the lower-level passenger

assignment, which is based on the congested assignment model of de Cea and Fernández (1993). The assignment is solved using a diagonalization approach.

Additional work has been done in consideration of special scheduling cases, particularly in high-demand corridors. These include short-turning, zone scheduling (or express services), and deadheading. In short-turning, some vehicles on a route will serve only one segment of the route before returning to the terminal. The objective is to reduce the total number of vehicles serving a route, while still meeting passenger demand and/or minimum levels of passenger service. Furth (1987) presented a model to consider short-turn design in which the objective is to minimize the total fleet size serving a route, with constraints that the load cannot exceed capacity at any point on either the full route or on the short-turn segment. For a given short-turn segment, the problem is cast separately for different multiples of the full route: a 1:1 strategy implies one vehicle on the full route for every one on the short-turn segment; a 1:2 implies one on the full route for every two on the short-turn route, etc. This problem is formulated and solved as a linear program with two decision variables: the frequency on the full route; and the relative offset of the dispatch times on the short-turn route versus the full route. The offset is used to balance the loads on the full route and the short-turn route, depending on the loading pattern over the common route segment. From the continuous linear program solution, a simple rounding technique can be used to find offsets at the nearest minute. Additional deadheading and interlining options are considered in a heuristic technique to reduce the total fleet requirement.

The problem description by Ceder (1989) considered a variety of possible short-turn segments on a route. As an input, a full route schedule is assumed. With this information, Ceder (1989) presented a method to determine which trip segments in the schedule could be eliminated by including short-turn trips. This uses a heuristic to minimize the maximum headway for a route segment when the short-turn trip is introduced. Once the short-turn trips are scheduled, a new estimate of the fleet size is found using the technique of Stern and Ceder (1983), based on the creation of deadheading trips and interlining of trips. Since it works with an existing (initial) route schedule, the technique results in both the definition on the short-turn segments as well as the schedule of the short-turn trips.

A more recent investigation by Site and Filippi (1998) posed the short-turning problem as one of determining the short-turn segment, the types of vehicles to operate on both the full route and the short-turn segment, and the frequency of service on both the full route and the short-turn segment. The objective function is to maximize the net benefits, given as the passengers' consumer surplus less the net subsidy (total costs minus fare revenues) for the operator. Costs for the operator include both capital and operating costs. In this model, passenger demand is endogenously determined as a function of the selected frequency. The model includes constraints to ensure demand does not exceed capacity, and a constraint on the maximum available subsidy. The problem is formulated as a nonlinear program. Site and Filippi (1998) decomposed

the full problem into smaller subproblems; each subproblem is solved for the optimal frequencies, for a given short-turn segment and set of vehicle types. These subproblems are solved heuristically using random search methods, and the global solution is found as the subproblem that maximizes the objective of net benefits.

The zone scheduling (or express service) problem was introduced by Jordan and Turnquist (1979) as a strategy to improve service reliability on bus routes. In the zone scheduling concept, a bus serves only selected segments of a route as it travels from one terminus to the other. In their simplified route model, Jordan and Turnquist (1979) assume that all passengers are destined for a single terminus. Their decision variables are the number of zones, the first stop in each zone, and the number of buses allocated to serve each zone. The problem is formulated as one of minimizing the passenger utility, comprising the expected value and variance of the waiting time and on-board travel time for all passengers. The problem is formulated using a dynamic programming recursion, in which the stages are the number of zones and the state variables are the combination of beginning stop and the number of buses allocated to that zone. The means and variances of waiting times and on-board running times are formulated using a stochastic model of transit service calibrated from data in Chicago.

The work of Furth (1986) extended the model of Jordan and Turnquist (1979) for several additional cases: (1) zone scheduling where additional stops outside the zone can be served for alighting (inbound) and for boarding (outbound); (2) zone scheduling where the zone boundaries are asymmetric, depending on the direction of the trip during any given period; and (3) zone scheduling for branching corridors. Again, dynamic programming is used to solve these cases of the zone scheduling problem.

Finally, Furth (1985) presented a model for deadheading, in which the desirable headway in the off-peak direction is higher than that in the peak direction. In providing service in the off-peak direction, vehicles not in service are deadhead back to the initial terminal in the peak direction. This strategy, while offering less service in the off-peak direction, may allow sufficient time savings for fleet size reduction. The problem of determining the minimum fleet size, for given (fixed) maximum headways in both the peak and off-peak direction, is found by solving a maximum flow problem on a time-space network. When the headway constraints are changed to inequalities (i.e., with only a maximum headway), it is possible that shorter headways could yield fleet size reductions, due to the scheduling requirements for these trips. Furth (1985) presented an algorithm to solve for the minimum number of vehicles, using the ratio of the number of peak trips per outbound trip in service. Related techniques are used to solve for the optimal headways in the peak and off-peak directions for a given fleet size, for two cases: (1) minimizing the total passenger waiting time; and (2) minimizing a combination of passenger waiting time and operating costs.

3.2 Timetabling

Timetabling is the process of converting the desired frequency of service on each fixed route into a schedule. The inputs to this process include the route structure, including running times between major timepoints, the frequency of service, and any necessary layover times at terminals or schedule slack (extra time built into the schedule) at the major timepoints on the route. The result is a set of trips and the scheduled times at the terminals and major timepoints on the route.

In most treatments of transit planning methods, timetabling is included within operational planning, since it occurs frequently, with every service adjustment (e.g., every 3–6 months). Also, it is from the timetables that vehicle and crew schedules are constructed. In this chapter, it is included as an element of tactical planning in the sense that it involves determining the schedule to maximize passenger service. This is in contrast to the vehicle and crew scheduling problems that are typically associated with operations planning, which are intended to minimize transit operating costs.

In many cases, the timetabling is relatively straightforward (Ceder, 1986), primarily working on the assumption that headways are constant and that demand is relatively uniform over the time period of interest. Some clock time is specified for the first vehicle trip of each period, and vehicle departures from a terminal or a maximum load point are set as multiples of the desired headway. Estimated running times between timepoints are used to determine the schedule, and layover times are used to schedule return trips. The simplicity of this task may explain in part why it has received relatively little attention from researchers. In this section, some methods for timetabling are described. However, a major complication in timetabling occurs when schedules are intended to be coordinated at a transfer stop or terminal; methods to create timetables under these circumstances are also presented in this section.

Initial work into the timetabling problem for time-dependent passenger arrivals was suggested by Newell (1971), Salzborn (1972), and Hurdle (1973a, 1973b). These works formulated the timetable problem for a single route with the objective of minimizing passenger waiting time. These results suggest that the optimal rate at which vehicles are dispatched is proportional to the square root of the passenger arrival rate, but with the constraint that vehicle capacity cannot be exceeded. These models were extended by Sheffi and Sugiyama (1982) to consider multiple origin–destination pairs (complicating the derivation of the capacity constraint) and to the case of boarding-dependent dwell times.

The work of Wirasinghe and Liu (1995) considers the problem of determining optimal schedule slack times at intermediate timepoints on a route. Given the running time distributions on the route and a schedule-based holding policy, the research gives a model for determining the slack time by minimizing the sum of the passenger waiting time, the passenger schedule delay, and the

operating cost. The problem is solved using first-order conditions on the objective function.

Perhaps the most pressing challenge in timetabling is the synchronization of vehicle timetables so that transfers within the network are well timed. Specifically, one would like to time the arrival of a vehicle on one route with that on another route so that passengers transferring between routes can make the connection with the minimum waiting time. Much of the early work on this problem focused on methods for synchronization at a single timepoint; more recent methods have used heuristics for larger network problems. However, the combinatorial nature of the problem indicates that it is NP-hard, and the computational issues of exact solutions are still vexing. This limitation is important to note, in that only the more recent approaches have considered more practical problems.

The work of Salzborn (1980) considers two related problems. The first problem is determining the feasibility of scheduling a single transfer route through a series of transfer locations ("interchanges"). Inputs to this analysis include the running times on the transfer route, the slack time built into the transfer route and all connecting services ("feeder routes") at these interchanges, and the minimum time required for passengers to make the transfer. In this analysis, the feasibility of the schedule for the transfer route depends on the ability of the schedule to meet the necessary time windows at each interchange. For the second problem, Salzborn (1980) considers the scheduling of the feeder routes at the interchange, and derives conditions under which a feasible feeder route schedules can be constructed. In this case, the feasible scheduling of the feeder routes requires that the headway on these routes be a multiple of the headway on the transfer route. In addition, the scheduling of the feeder routes requires that departures and arrivals at the interchange be balanced (the total number of departures equals the total number of arrivals) over one headway on the feeder routes, so that time slots at the interchange can be scheduled.

Hall (1985) considers the more specific problem of scheduling at the interchange when the feeder route may be delayed. Under an assumed exponential distribution of this delay, Hall (1985) derives equations for the optimal slack time, based on the objective of minimizing passenger delay. In this case, "slack time" is defined as the time between the scheduled arrival on the feeder route and the scheduled departure on the transfer route. A similar approach was used by Knoppers and Muller (1995) to characterize the optimal slack time on the transfer route, with a normal distribution of arrival times on the feeder route. Knoppers and Muller (1995) also investigate the possible reductions in average waiting time if a holding policy is used: a vehicle on the transfer route is held until the feeder bus arrives. In this case, the optimal slack time can be reduced, with corresponding reductions in the average waiting time. However, the work does not examine the implications of additional waiting at downstream stops from the holding strategy.

There has also been a number of more analytic studies to optimize both the slack time and the headways of connecting routes at a transfer point. Purely

analytic optimization models for a single transfer point have been developed by a number of researchers, including Lee and Schonfeld (1991) and Chien and Schonfeld (1998). More recently, these analytic models have been extended to cover multiple transfer points and multiple modes. Heuristics for this problem, to solve for the slack times and any common headways among routes, have been proposed by Chowdhury and Chien (2002) and Ting and Schonfeld (2005). Both of these studies use a combination of operator costs, from the added slack time, and user costs, including waiting time, transfer time, and in-vehicle time.

A more rigorous treatment of the transfer synchronization problem was presented originally by Klemt and Stemme (1988) for a completely deterministic problem, and later by Bookbinder and Désilets (1992) for the case of random delay. The synchronization problem is defined as one to determine the ideal "offsets" for the schedule for each route r in the set R. Here, an *offset* t_r for a given route r is defined as the minutes after some given time at which the first departure from the terminal is scheduled. If the headway on route r is h_r, the possible offsets are assumed to be in the set of integers $T_r = \{0, 1, \ldots, h_r - 1\}$. Suppose the set of transfer opportunities between routes is given as the set $K = \{1, 2, \ldots, k\}$, with the set A_{ij} describing the complementary set of pairs of routes i and j representing transfer opportunity k (i.e., at the intersection of routes i and j). The utility of a given transfer opportunity is given as $D_k(t_i, t_j)$, and the number of passengers making this transfer is given as n_k. Then, the optimization problem to determine the optimal offsets is given as

$$\text{minimize} \quad \sum_{i \in R} \sum_{j \in R} \sum_{k \in A_{ij}} n_k D_k(t_i, t_j) \tag{24}$$

$$\text{subject to:} \quad t_i \in T_i, \quad \forall i \in R. \tag{25}$$

To solve this model, Bookbinder and Désilets (1992) use a heuristic developed by Rapp and Gehner (1976), in which the offset of each route is determined iteratively. In Bookbinder and Désilets (1992), several utility functions $D_k(t_i, t_j)$ based on the passenger waiting time are evaluated, using a simulation model to generate random transfer arrivals. These are evaluated initially for a single transfer connection and also two small networks with multiple transfer connections.

Ceder and Tal (1999) and Ceder et al. (2001) introduced a model to maximize the number of synchronized connections between routes. The objective is simply to maximize the number of potential connections that can be made. In this formulation, the decision variables include the offset of the initial trip in the schedule (as before), and the headway of each vehicle trip is permitted to vary from some minimum to a maximum headway. This creates more flexibility in the construction of schedules, as each vehicle on the route may have a different headway. The problem is formulated as a mixed integer linear program, and solved using a heuristic that processes the nodes sequentially, using some selection criteria. For each node, the headways of all routes at the node are

matched if possible, and offsets of all connecting routes are set so as to create a simultaneous arrival of all routes at the given node. The heuristic proceeds through the set of interchanges until all vehicle departure times are set.

4 Operational planning

Given a set of timetabled trips to be operated, a set of available resources (buses and drivers) and their distribution among the transit agency depots, the operational planning phase aims at constructing vehicle and crew schedules that minimize total costs while respecting all operational constraints and work regulations. This phase also includes planning the assignment of the buses to parking slots in garages and their dispatch to bus schedules, as well as establishing bus maintenance schedules.

As mentioned in the Introduction, these various problems are usually solved sequentially. Figure 1 illustrates the usual solution sequence and the possible types of feedback. The frequency setting and timetabling tactical problems define the main input for the operational planning phase, namely the timetable. After this, the vehicle scheduling problem is solved first. This stage is crucial to assess whether the proposed timetable can be operated with the available fleet of vehicles and how costly it will be. When the vehicle scheduling problem is infeasible or its operating cost is excessive, a revision of the timetable and possibly the route frequencies is performed in order to facilitate vehicle scheduling. Once a vehicle schedule is determined, driver duty scheduling, which consists of constructing anonymous work days for the drivers, is performed to ensure a complete coverage of the vehicle schedule at a reasonable cost. When this problem is not feasible or too costly, the vehicle schedule must

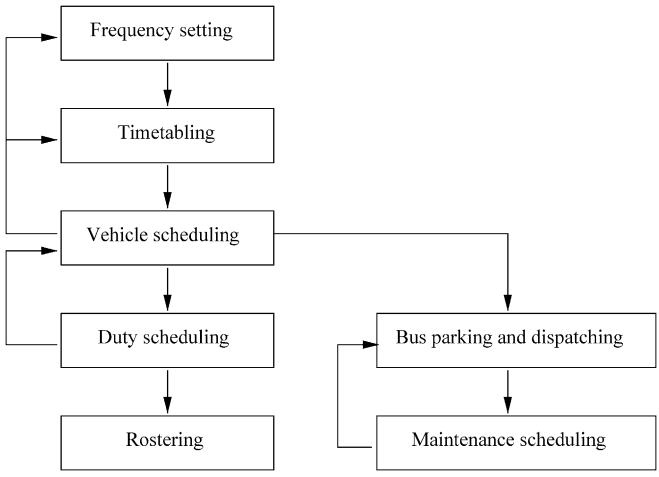


Fig. 1. Relationships between the tactical and operational problems.

be updated to ease the construction of the driver duties. Once the driver duties are known, a rostering problem is solved to establish the personalized driver schedules over a given time horizon (e.g., monthly). This stage rarely requires one to revise the previous decisions since part-time drivers are usually available to provide additional flexibility in the rosters. In parallel to the driver scheduling problems, the bus parking and dispatching problem as well as the bus maintenance scheduling problem are tackled once the vehicle schedule is known. These two problems are more or less solved on a daily basis since they are heavily impacted by the perturbations of the planned vehicle schedule. These two problems are also quite dependent.

Looking at Figure 1, one can see that feedback may frequently occur during the operational planning process. This shows that there is an opportunity for integrating some of these steps. In this regard, Section 4.3 discusses the integration of vehicle and duty scheduling.

4.1 Vehicle scheduling

Vehicle scheduling plays an important role in the management of a public transit agency since it is the first planning step where the primary focus is put on minimizing costs, namely, the acquisition, and operational costs of the buses. Previous steps put a large emphasis on passenger service, which is fixed for vehicle scheduling. Indeed, at this stage, the service offered to the customers is completely determined by the fixed trip timetable. In general, the same daily timetable applies for the weekdays, while a different timetable is defined for days on the weekend. These timetables are usually valid for a certain period of time (a season). Thus vehicle scheduling at a planning level needs to be performed once per timetable and season. It should be noted that in large cities buses operate 24 hours per day, which makes it more difficult to define a daily problem. However, given the low volume of activities during night-time, a 24-hour timetable is split into a day timetable and a night timetable in practice.

The vehicle scheduling problem faced by the public transit agencies corresponds to the single-depot vehicle scheduling problem (SDVSP) when the agency operates its fleet of buses out of a single depot, and to the multidepot vehicle scheduling problem (MDVSP) when several depots are used or when several vehicle types are available. This problem can be stated as follows. Let $\mathcal{T} = \{1, 2, \ldots, n\}$ be a set of n timetabled trips where trip $i \in \mathcal{T}$ starts at time s_i and ends at time e_i. These trips are qualified as *active* since passengers travel along them. Denote by τ_{ij} the travel time (possibly including some layover time) between the end location of trip i and the start location of trip j. We assume that this travel time is the same for all vehicles. Two trips i and j are said to be *compatible* if and only if they can be covered consecutively by the same vehicle (j immediately follows i), that is, if and only if $e_i + \tau_{ij} \leqslant s_j$. The traveling between two such trips is called a *deadhead trip* since there are no passengers on board.

Let $K = \{n+1, n+2, \ldots, n+m\}$ be the set of m depots housing the buses that must be assigned to cover the active trips. Depot $k \in K$ manages v^k identical buses which must start and end their schedule at this depot. A bus leaving a depot to reach the start location of an active trip is said to be performing a *pull-out trip*, while it performs a *pull-in trip* when it returns to the depot from the end location of an active trip. A *feasible schedule* for a bus housed in depot k is composed of a pull-out trip starting at k, a sequence of active trips separated by deadhead trips, and a *pull-in trip* ending at k. Consecutive active trips must be pairwise compatible.

The cost structure is as follows. A cost is incurred each time that a vehicle performs a deadhead, pull-out or pull-in trip. This cost is denoted by c_{ij} for the deadhead trip connecting trips i to j, by c_{kj} for the pull-out trip linking depot k to the start location of trip j, and c_{ik} for the pull-in trip returning to depot k from the end location of trip i. Note that the active trips bear no cost since they represent a fixed amount for any feasible solution. Note also that vehicle fixed costs can be added to the pull-out or the pull-in trip costs. The cost of a schedule is simply the sum of the costs of the trips it contains.

The MDVSP can be defined as the problem of finding a set of feasible vehicle schedules such that each active trip $i \in \mathcal{T}$ is covered by exactly one schedule, at most v^k schedules are defined for each depot $k \in K$, and the sum of the schedule costs is minimized. The SDVSP simply corresponds to the case where $|K| = 1$. In certain versions of the MDVSP additional constraints are considered. For instance, there may exist trip–depot compatibility constraints which restrict the set of depots that can provide a vehicle to perform an active trip, especially when the depots are associated with different vehicle types (see Costa et al., 1995; Löbel, 1998). A soft version of these constraints, when they take the form of preferences rather than strict constraints, can also be handled by defining depot-dependent deadhead costs. Another example of additional constraints consists of imposing a maximum duration or length to every vehicle schedule (Freling and Paixão, 1995). Such a constraint may be needed in a extra-urban context where the potential for driver exchanges is restricted or when fueling considerations must be taken into account.

It should be noted that, in the context of public transit, a deadhead trip that involves a long waiting time before the start of the next active trip is often replaced by a pull-in trip, an idle period at the depot, and a pull-out trip. The vehicle schedules are then seen as sequences of *vehicle blocks*, where each block consists of a sequence of trips that starts and ends at the same depot without returning to it in the middle of the sequence.

The SDVSP arises for small to medium-size transit agencies that rely on a single depot. It can also appear as a subproblem of the MDVSP. The SDVSP is solvable in polynomial time. In fact, it can be modeled as a minimum-cost network flow problem. It has also been formulated as a linear assignment problem, a transportation problem, a quasiassignment problem, and a matching problem. Surveys on the SDVSP and its extensions can be found in Daduna and Paixão (1995) and Desrosiers et al. (1995). Recently, Freling et al. (2001b)

proposed an efficient auction algorithm for solving the quasiassignment formulation of the SDVSP. Based on this algorithm, they also developed a two-phase approach where blocks are built first and combined afterwards, and a core-oriented approach that starts with a network containing a subset of the arcs (the core) and adjusts it iteratively according to the reduced cost of the arcs not considered. The two-phase approach is only valid under certain cost assumptions. Computational experiments showed that these approaches outperform the algorithms previously exposed in the literature.

The MDVSP is common in medium-size transit agencies, and inevitable in larger ones. As proposed by Ribeiro and Soumis (1994), it can be modeled using an integer multicommodity flow formulation as follows. Associate with each depot $k \in K$ a directed graph $G^k = (N^k, A^k)$, where N^k and A^k denote its sets of nodes and arcs, respectively. The node set is defined by $N^k = \mathcal{T} \cup \{k\}$. The arc set is given by $A^k = C \cup (\{k\} \times \mathcal{T}) \cup (\mathcal{T} \times \{k\})$, where C is a subset of $\mathcal{T} \times \mathcal{T}$ that contains an arc $(i, j) \in \mathcal{T} \times \mathcal{T}$ if and only if trips i and j are compatible. Then, define a binary variable X_{ij}^k for each $k \in K$ and each arc $(i, j) \in A^k$ that indicates the flow (0 or 1) of buses originating from depot k on the arc (i, j).

Using this notation Ribeiro and Soumis (1994) formulated the MDVSP as follows:

$$\text{minimize} \quad \sum_{k \in K} \sum_{(i,j) \in A^k} c_{ij} X_{ij}^k \tag{26}$$

subject to:

$$\sum_{k \in K} \sum_{i:(i,j) \in A^k} X_{ij}^k = 1, \quad \forall j \in \mathcal{T}, \tag{27}$$

$$\sum_{j \in \mathcal{T}} X_{k,j}^k \leqslant v^k, \quad \forall k \in K, \tag{28}$$

$$\sum_{i:(i,j) \in A^k} X_{ij}^k - \sum_{i:(j,i) \in A^k} X_{ji}^k = 0, \quad \forall k \in K, j \in \mathcal{T} \cup \{k\}, \tag{29}$$

$$X_{ij}^k \in \{0, 1\}, \quad \forall k \in K, (i, j) \in A^k. \tag{30}$$

The objective function (26) aims at minimizing total costs. Constraints (27) ensure that each active trip is covered exactly once, while constraints (28) limit the number of buses that can be used from each depot. Flow conservation and binary constraints are given by (29) and (30), respectively.

The MDVSP has been studied for more than twenty-five years. Given that it is an NP-hard problem when $m \geqslant 2$ (Bertossi et al., 1987), the early work on this problem focused on heuristic algorithms (for reviews on these methods, see Dell'Amico et al., 1993; Odoni et al., 1994). Since the end of the 1980s, several exact algorithms have been proposed in the literature: Carpaneto et al. (1989), Ribeiro and Soumis (1994), Forbes et al. (1994), Bianco et al. (1994),

Löbel (1998), Mesquita and Paixão (1999), Hadjar et al. (2006), and Kliewer et al. (2006). The results in the last three papers clearly show that real-world large-scale instances can be solved efficiently. In the following we present the methodologies introduced in these three papers.

The approach proposed by Löbel (1998) consists of solving the linear relaxation of the multicommodity network flow model (26)–(30) using a column generation method directly on this formulation; that is, the generated variables are the X_{ij}^k. Before starting the column generation process, a heuristic procedure is used to find a feasible solution. The positive-valued variables of this solution are added to the initial restricted master problem in order to speed up the solution process. Then, at each column generation iteration, the restricted master problem is solved by the dual simplex algorithm and the columns are generated based on a so-called Lagrangian pricing strategy (discussed below) and the standard reduced cost criterion. When the progress in the objective value becomes too small, only the standard reduced cost criterion is used and the restricted master problem is re-optimized by the primal simplex algorithm. Using this methodology, the optimal linear relaxation solution found by Löbel (1998) was already integer for most of the instances treated. When this was not the case, a simple rounding procedure was used to derive an integer solution that was often proved to be optimal.

Lagrangian pricing is a strategy which allows the simultaneous generation of negative reduced cost variables and nonnegative reduced cost variables that complement well the former set of variables. In this way, the column generation process does not require additional iterations to identify these complementary variables as is often the case with the traditional pricing strategy. Given a Lagrangian relaxation of the linear relaxation of the model (26)–(30), the Lagrangian pricing proposed by Löbel (1998) consists of solving the Lagrangian subproblem for a given set of multipliers, namely the values of the dual variables associated with constraints (27)–(29) in the current restricted master problem. All the variables taking a positive value in the solution of this subproblem are then candidates that can be added to the restricted master problem, even if their reduced costs are nonnegative.

Löbel (1998) considered at each iteration two Lagrangian relaxations. In the first, the trip covering constraints (27) are relaxed in the objective function to obtain m independent minimum-cost flow problems that are easily solvable. For the second Lagrangian relaxation, the redundant covering constraints

$$\sum_{k \in K} \sum_{j:(i,j) \in A^k} X_{ij}^k = 1, \quad \forall i \in \mathcal{T}, \tag{31}$$

are added to the formulation (26)–(30) before relaxing constraint sets (28) and (29). The resulting Lagrangian subproblem is also a minimum-cost flow problem that can be solved by inspection. Its solution can however produce bus schedules containing deadhead trips assigned to different depots.

Using this column generation approach, Löbel (1998) reports solving real-world instances from German public transit companies involving up to 49 de-

pots and 24,906 trips. It should be noted however that trip–depot compatibility constraints were considered, yielding a maximum average of 4 depots per trip among these large instances.

Recently, Hadjar et al. (2006) developed a branch-and-bound approach for the MDVSP that combines column generation, variable fixing, and cutting planes. As introduced in Ribeiro and Soumis (1994), traditional column generation is used to compute a lower bound at each node of the branch-and-bound search tree. This column generation process is executed on a set partitioning type reformulation of the MDVSP that can be derived from model (26)–(30) by applying the Dantzig–Wolfe decomposition principle (Dantzig and Wolfe, 1960). In contrast to Löbel (1998) approach, columns in this set partitioning model correspond to vehicle schedules. They are generated by solving shortest path problems.

The variable fixing strategy used by Hadjar et al. (2006) is similar to the one developed by Bianco et al. (1994) and consists of fixing to zero the variables X_{ij}^k that satisfy the following criterion. To simplify notation, we rewrite model (26)–(30) as $\min\{cx \mid Ax = b, x \in \mathbb{Z}_+^\eta\}$, where $x = (x_i)_{i=1}^\eta$ is a vector of η $(= \sum_{k \in K} |A^k|)$ variables, $\mathbb{Z}_+$ is the set of nonnegative integers, and the equality $Ax = b$ is, in fact, an inequality $(Ax \leqslant b)$ for constraint set (28). Denoting by $\bar{x}$ a feasible solution to this problem and by $\bar{\pi}$ a feasible solution to the dual of its linear relaxation, a variable x_i can be set to zero if its reduced cost is greater than or equal to $c\bar{x} - \bar{\pi}b$. Hadjar et al. (2006) computed a first feasible solution $\bar{x}$ by performing a depth-first search without backtracking in the branch-and-bound tree and imposing multiple decisions at each branching node. At each node of the branching tree, part of the dual solution $\bar{\pi}$ is provided by the dual solution produced by the column generation method at this node and the remainder is found by solving shortest path problems. This variable fixing strategy, which is performed at each node of the tree, can fix over 90% of the X_{ij}^k variables in most instances treated by Hadjar et al. (2006).

Hadjar et al. (2006) proposed to add at each node of the search tree cutting planes that are related to the odd cycles in the MDVSP underlying network. These valid inequalities are lifted through a heuristic procedure. The authors showed that, under certain conditions, the lifted inequalities define facets of the convex hull of the feasible solution set. With this approach, Hadjar et al. (2006) succeeded in solving randomly generated MDVSP instances that involve up to 6 depots and 750 trips. It should, however, be mentioned that these results are difficult to compare with the results obtained by Löbel (1998) or Kliewer et al. (2006) (see below) because the characteristics of the test problems differ significantly from one paper to the other.

Instead of using model (26)–(30), Kliewer et al. (2006) developed a multicommodity network flow model based on a time-space network. In fact, when several trips start and end at the same terminus, a substantial reduction in the number of variables can be obtained by using a sequence of waiting arcs at each terminus instead of representing explicitly all possible connections. Kliewer

et al. (2006) also applied an aggregation procedure for reducing the number of arcs representing potential deadhead trips. The resulting model is solved to optimality with the CPLEX MIP solver. With this approach, the authors report solving large real-world instances. In particular, they solved one instance that involves 7068 trips, 5 depots, and 124 termini in approximately 3 hours of computational time.

An interesting extension to the MDVSP is the possibility of changing the scheduled departure times of the trips within certain time intervals, called *time windows*. Such flexibility on departure times, which can be considered when the frequency on a route is not too high, can often yield significant savings by providing additional possible deadhead trips that would be infeasible otherwise. To our knowledge, this extension has been tackled first by Mingozzi et al. (1995) who adapted the methodology they have developed for the MDVSP in Bianco et al. (1994). Their approach consists of solving by branch-and-bound a set partitioning model that contains a reduced set of columns. The reduction is obtained by variable fixing as described above. More recently, Desaulniers et al. (1998b) have proposed a branch-and-price approach for this extension that generalizes the work of Ribeiro and Soumis (1994). They also showed that, with a slight modification, this approach is capable of handling an exact cost on the waiting time occurring between two consecutive trips. Given the time windows, such waiting times and their ensuing waiting costs cannot be computed a priori; they must be computed during the solution process.

4.2 Duty scheduling

Duty scheduling, also known as driver scheduling, is the second step in the process of planning the operations for a public transit agency. As with the vehicle scheduling problem, the duty scheduling problem (DSP) is important from an economic point of view since it determines most of the wages paid to the drivers. The DSP is separable by depot and consists of determining the work days (also called *duties*) of the drivers based at a depot in order to cover all the vehicle blocks assigned to this depot. Since a driver exchange can occur at various points along a vehicle block, all blocks are divided into a sequence of *segments* according to these *relief points*. The consecutive segments along a block assigned to the same driver are collectively called a *piece of work*. Duties are therefore composed of pieces of work that are usually separated by breaks. Different duty types, differing, for instance, by the number of pieces of work they can contain and their possible starting times and durations, can be considered. As examples, there may exist straight duties that contain a single piece of work, and split duties containing two pieces of work. Duties are subject to a wide variety of safety regulations and collective agreement rules such as a maximum duty spread, a maximum duration of a piece of work, and a predefined time interval in which a break must be awarded. These rules vary according to the duty type.

In general, the objective of the DSP is twofold and consists of minimizing first the total number of duties and second the total number of worked hours. Using duty fixed costs and an hourly rate for the worked hours, this objective is usually transformed into one that minimizes total cost. In summary, the DSP can be stated as follows. Given the segments of a set of vehicle blocks, find a set of valid duties that covers all these segments and minimizes total cost. Additional constraints such as a limit on the number of duties of a certain type can also be taken into account.

The DSP can be formulated as a set partitioning problem that relies on the following notation. Let S be the set of block segments to cover and D the set of valid duties. Denote by c_d the cost of duty d (including fixed costs and wages) and by a_d^s a binary parameter that takes value 1 if duty d covers segment s, and 0 otherwise. Finally, define a binary variable Y_d for each duty $d \in D$ that indicates if duty d is retained in the solution. Using this notation the DSP is formulated as

$$\text{minimize} \quad \sum_{d \in D} c_d Y_d \tag{32}$$

subject to:

$$\sum_{d \in D} a_d^s Y_d = 1, \quad \forall s \in S, \tag{33}$$

$$Y_d \in \{0, 1\}, \quad \forall d \in D. \tag{34}$$

The objective function (32) consists of minimizing total duty costs. Constraint set (33) ensures that a driver is assigned to each block segment. Finally, binary requirements on the Y_d variables are expressed by (34). It should be noted that all work rules defining the validity of the duties are taking into account in the definition of set D. In some cases, this set or part of it can be enumerated a priori using an enumeration algorithm that considers these rules (for instance, see Smith and Wren, 1988). Otherwise, it can be defined implicitly as a set of constrained paths in one or several networks, where the constraints are used to model the complex work rules (for instance, see Desrochers and Soumis, 1989).

Several authors formulate the DSP as a set covering model that allows the over-covering of each segment (the equalities in (33) are replaced by greater-than-or-equal-to inequalities). This over-covering is usually not acceptable, but solving this model often produces a solution that contains very little or no over-covering at all, especially when assigning one driver to a segment is cheaper than assigning several drivers to it. In this case, over-covering can usually be easily eliminated a posteriori using a heuristic procedure. The main advantage of a set covering model over a set partitioning model is its flexibility which allows more rapid computation of a feasible continuous solution and good heuristic integer solutions.

The DSP, modeled as a set partitioning or a set covering problem, is much more difficult to solve than the MDVSP due to the complexity of the work

rules and the huge number of duties that are valid in real-world DSP instances. Indeed, since some of these rules can only be modeled using nonlinear relationships, one has to rely on a formulation similar to (32)–(34) to avoid explicit nonlinear constraints. This formulation however contains a huge number of variables in practice, making it very difficult to solve to optimality.

As surveyed in Wren and Rousseau (1995), several heuristic approaches were proposed before the 1990s. One of the most successful consists essentially of generating a priori a subset of the valid duties and solving a set partitioning/covering model (32)–(34) restricted to this subset (for instance, see Smith and Wren, 1988). The subset of valid duties is composed of promising duties that offer various possibilities for covering each block segment. The restricted model is generally solved by a heuristic integer linear programming method.

Research on this type of approach is still ongoing. In 1999, an attempt was made by Curtis et al. (1999) to solve the restricted set partitioning model using a hybrid constraint programming/linear programming heuristic method where the linear programming solutions are used to guide variable and value ordering in the constraint programming algorithm. One nice feature of this methodology is that its constraint programming component is capable of handling nonlinear constraints which can arise from certain work rules. However, the authors could not solve instances involving more than 203 segments and 26 duties. More recently, Fores et al. (2002) have incorporated a column generation strategy to the solution process of the restricted set covering model. Column generation, which is applied only at the root node of the branch-and-bound search tree, consists of generating as needed negative reduced cost columns from a superset of a priori enumerated valid duties. This novelty improves the quality of the solutions while slightly increasing solution times. For instance, they can solve medium-size instances involving close to 90 duties in one hour of computational time.

Column generation for the DSP was introduced by Desrochers and Soumis (1989). In their approach the master problem corresponds to the linear relaxation of the set covering model. The subproblems are constrained shortest path problems which are solved by a generalized version of the dynamic programming algorithm of Desrochers and Soumis (1988) (see Desaulniers et al., 1998a; Irnich and Desaulniers, 2005). Integer solutions are found by a branch-and-bound scheme, where column generation is used at each node of the search tree to compute a lower bound. The overall approach obtains optimal solutions for small instances and near-optimal solutions for larger instances. In 1995, Desrosiers and Rousseau (1995) reported solving DSP instances involving 156 duties using a commercial version of this branch-and-price approach.

In the last fifteen years, branch-and-price approaches have been applied with success to a wide variety of vehicle routing and crew scheduling problems. When the size of these problems are huge as in real-world DSPs, accelerating strategies, various heuristics, and stabilization techniques, such as the ones reported in Barnhart et al. (1998), du Merle et al. (1999), and Desaulniers et al. (2002), must be included in these approaches to produce good quality solutions

in acceptable computational times. In a recent paper on the DSP, Borndörfer et al. (2003) pursued this direction by presenting a heuristic branch-and-price approach that calls upon several speed-up strategies. Firstly, instead of solving the master problem by the simplex or barrier algorithm, they proposed to solve its dual using a heuristic coordinate ascent method combined with a boxstep stabilization method. Secondly, the constrained shortest path subproblems are solved by an enumerative algorithm that relies on lower bounds computed by Lagrangian relaxation. Finally, they suggested a heuristic branching scheme without backtracking that fixes at each branching node a duty variable Y_d to one. This variable is selected from a set of twenty candidate variables using a probing strategy. For each candidate variable, this strategy fixes it to one and evaluates the impact of this decision on the master problem solution value without generating additional columns. The selected variable is the one yielding the smallest deterioration of the master problem solution value. After branching, when this deterioration is less than a predetermined threshold value, no columns are generated and variable fixing is performed again. Using this customized column generation approach, Borndörfer et al. (2003) solved large real-world DSP instances involving close to 2000 segments and over 110 duties.

Freling et al. (1999, 2003) proposed to solve the linear relaxation of the DSP by a column generation approach where the master problem is approximately solved by Lagrangian relaxation. At each column generation iteration, a constrained shortest path subproblem is solved in two main steps: first, pieces of work are generated by solving an all-pairs shortest path problem; and second, duties are generated from these pieces of work by solving a constrained shortest path problem. To avoid generating the same column twice, the Lagrange multipliers are modified before generating new columns in such a way that all columns already generated get a nonnegative reduced cost. Once the linear relaxation is solved, a set covering problem involving all columns generated along the way is heuristically solved to find a feasible integer solution. As reported in Freling et al. (2003), this approach can easily solve small-size DSP instances with 238 segments and 24 duties. No results on larger instances are reported, however, since the DSP was solved only for comparison with an integrated vehicle and crew scheduling approach that is discussed in the next section.

One drawback of the column generation approach that relies on constrained shortest path subproblems is the inability to model all work rules that may define the validity of a duty. One way of overcoming this drawback is to ignore these rules at the subproblem level and simply to reject all generated columns that violate them. Another way that was suggested by Borndörfer et al. (2003) consists of defining an infeasible path constraint for each identified illegal column and including these constraints in the subproblems. A third alternative has been proposed by de Silva (2001), who suggested formulating the subproblems as flexible constraint programming models and solving them using constraint programming tools. He succeeded in solving real-world DSP instances with complex work rules involving up to 495 segments.

Recently, Gintner et al. (2004) proposed a DSP approach that benefits from the fact that several vehicle schedules may be optimal. Indeed, given a feasible vehicle schedule, the segments derived from the blocks of this schedule can be rearranged into different blocks yielding the same vehicle costs. Their DSP model allows for this possibility by enumerating all feasible pieces of work and duties that can be obtained from the segments. Their approach solves the linear relaxation of this model using column generation and computes, using the CPLEX MIP solver, an integer solution for the restricted set covering model containing only the columns generated. This approach was tested on random datasets involving up to 400 trips with one segment per trip. These tests showed that substantial savings in the number of duties can be achieved from this additional flexibility.

Research investigating the use of metaheuristics for solving the DSP has also been carried out recently. Kwan et al. (1999, 2001) proposed a genetic algorithm that relies on the linear relaxation solution of a restricted set covering model to identify important traits that should appear in the optimal integer solution. Shen and Kwan (2001) developed a tabu search approach that involves multiple neighborhoods and an appropriate memory scheme. Lourenço et al. (2001) introduced a genetic and a tabu search algorithm to solve DSPs involving multiple objectives such as minimizing the number of duties, minimizing the number of duties with a single piece of work, minimizing the number of vehicle changes, and minimizing the over-covering when allowed. Both of these algorithms use for large instances a greedy randomized adaptive search procedure (GRASP) as an intensification tool. All these metaheuristics are fast and produce solutions that are comparable (in terms of quality) to the solutions produced by an approach based on a restricted set partitioning/covering model, similar to that of Smith and Wren (1988). For instance, Shen and Kwan (2001) reported solving a DSP involving 859 segments and 106 duties in less than 18 minutes with their tabu search algorithm.

4.3 Integrated vehicle and duty scheduling

In general, vehicle scheduling is performed before duty scheduling in the operational planning process of a public transit agency. Since driver relief opportunities are numerous in most contexts, an efficient duty schedule can often be obtained from a near-optimal bus schedule to yield an overall high-quality solution. On the other hand, when these relief opportunities are rare, as is the case in extra-urban mass transit systems or for a line-by-line scheduling process, a very efficient vehicle schedule may lead to a poor duty schedule or even to an infeasible DSP. Integrating vehicle scheduling and duty scheduling is therefore essential in these situations, and research on this topic has been conducted recently.

The integrated vehicle and duty scheduling problem (IVDSP) can be stated as follows. Given a set of timetabled trips and a fleet of vehicles assigned to several depots, find minimum-cost vehicle blocks and valid driver duties such

that each active trip is covered by one block, each active trip segment is covered by one duty, and each deadhead, pull-in, and pull-out trip (hereafter called an *inactive trip*) used in the vehicle schedule is also covered by one duty. As in the MDVSP, each block must start and end at the same depot and, as in the DSP, driver duties must comply with a set of work rules and each duty must be composed of trips that are covered by buses originating from the same depot. This last requirement is often mandatory since drivers are usually assigned to a depot. Additional constraints such as vehicle availability can also be imposed.

Next, we propose a formulation for the IVDSP that combines the models presented above for the MDVSP and the DSP. Besides the notation introduced in the previous two sections, the following notation is required. Let D^k be the set of valid duties for a driver assigned to depot k, and b_{dij} be a binary parameter equal to 1 if duty $d \in D^k$ covers the trip associated with arc $(i, j) \in A^k$ and to 0 otherwise. For each depot $k \in K$ and each duty $d \in D^k$, we define a binary variable Y_d^k that takes the value 1 if duty d is selected and the value 0 otherwise.

The proposed formulation for the IVDSP is as follows:

$$\text{minimize} \quad \sum_{k\in K} \sum_{(i,j)\in A^k} c_{ij} X_{ij}^k + \sum_{k\in K} \sum_{d\in D^k} c_d Y_d^k \tag{35}$$

subject to:

$$\sum_{k\in K} \sum_{i:(i,j)\in A^k} X_{ij}^k = 1, \quad \forall j \in \mathcal{T}, \tag{36}$$

$$\sum_{j\in\mathcal{T}} X_{k,j}^k \leqslant v^k, \quad \forall k \in K, \tag{37}$$

$$\sum_{i:(i,j)\in A^k} X_{ij}^k - \sum_{i:(j,i)\in A^k} X_{ji}^k = 0, \quad \forall k \in K, j \in \mathcal{T} \cup \{k\}, \tag{38}$$

$$X_{ij}^k \in \{0, 1\}, \quad \forall k \in K, (i, j) \in A^k, \tag{39}$$

$$\sum_{k\in K} \sum_{d\in D^k} a_d^s Y_d^k = 1, \quad \forall s \in S, \tag{40}$$

$$\sum_{d\in D^k} b_{dij} Y_d^k - X_{ij}^k = 0, \quad \forall k \in K, (i, j) \in A^k, \tag{41}$$

$$Y_d^k \in \{0, 1\}, \quad \forall k \in K, d \in D^k. \tag{42}$$

The objective (35) minimizes the sum of the vehicle and duty costs. Constraints (36)–(39) define the vehicle scheduling problem. They are identical to (27)–(30). Constraint sets (40) and (42) are the counterparts of (33) and (34) for the multidepot case. Finally, constraints (41) establish the link between the vehicle schedule and the duty schedule; that is, each inactive trip covered by a bus must also be covered by a duty assigned to the depot from which this bus originates.

In comparison with the DSP, the IVDSP is highly combinatorial since the inactive trips are unknown a priori; they have to be determined by the optimization process. In consequence, the number of possible valid duties is very large especially when multiple depots are considered. Given its complexity and its lesser importance, the IVDSP has not been addressed in the literature as much as the MDVSP and the DSP. Indeed, as surveyed in Freling et al. (1999), only a few heuristic approaches have been proposed in the 1980s and the early 1990s. However, it seems that this problem has lately attracted the attention of several researchers who developed solution approaches based on mathematical programming decomposition techniques.

Freling et al. (1999, 2003) addressed the single-depot IVDSP where no bus availability constraints are considered but the main objective consists of minimizing the overall number of buses and duties required to cover all active trips. Similar to the approach they proposed for the DSP, they developed a column generation approach where the master problem is solved by Lagrangian relaxation. In this case, all constraints involving duty variables are relaxed in the Lagrangian function, yielding a Lagrangian subproblem that corresponds to pricing out the duty variables and solving a single-depot vehicle scheduling problem. Thus, a feasible vehicle schedule is computed each time that the Lagrangian subproblem is solved. When the linear relaxation of the IVDSP is satisfactorily solved using this process, the last computed vehicle schedule is kept and used to define a DSP that is solved by their DSP column generation approach (see Section 4.2) to derive a feasible duty schedule. In Freling et al. (2003), the authors report solving real-world IDVSP instances involving up to 148 segments and 23 duties in reasonable computation times. Their results also show that small gains in the total number of buses and duties can be attained by solving the IVDSP instead of solving the vehicle scheduling problem and the duty scheduling problem sequentially. These gains are more substantial when drivers are not allowed to change buses after a break (see also Freling et al., 2001a).

In 2001, Haase et al. (2001) introduced a formulation that only involves duty variables and one bus counter variable which is used to apply a fixed cost per bus. This model can be partially derived from model (35)–(42), adapted to the single-depot case and without availability constraints, by substituting the X variables according to their definition in constraint set (41). Bus-count constraints, similar to the plane-count constraints of Klabjan et al. (2002), are added to complete the model. These constraints provide lower bounds on the number of buses required at specific times of the horizon, namely each time that a bus can leave the depot to reach just in time the beginning location of an active trip. Solving this model provides optimal duties and ensures that an optimal vehicle schedule can be obtained a posteriori using a simple polynomial-time procedure. To do so, Haase et al. (2001) proposed a branch-and-price approach that relies on several accelerating strategies such as dynamically generating the bus-count constraints and reducing the average number of nonzero elements in the constraint coefficient matrix by an appropriate constraint sub-

stitution. Two versions of this approach are presented: an exact version where branching is performed at the subproblem level, and a heuristic version where multiple branching decisions on the duty variables are made at every branching node. With the exact version of the algorithm, randomly generated IVDSP instances involving up to 400 segments and 60 duties were solved in less than 3 hours, while the heuristic version succeeded in solving instances with 700 segments and 121 duties within the same time frame.

Recently, Elhallaoui et al. (2005) developed a dynamic constraint aggregation algorithm for speeding up the solution process of set partitioning type problems solved by a column generation approach. This exact algorithm aggregates and disaggregates, as needed, the set partitioning constraints in order to reduce the size of the master problem and degeneracy. They tested this new approach on the single-depot IVDSP instances of Haase et al. (2001). They report reducing the time needed for solving the linear relaxation by up to 80% on instances involving up to 1280 segments. Furthermore, they observed that the number of fractional-valued variables in a linear relaxation solution decreases considerably with this methodology, yielding high expectations to compute rapidly optimal integer solutions.

The multidepot version of the IVDSP has been investigated in Huisman et al. (2005), where the authors presented two formulations for this problem that are generalizations of the single-depot models developed in Freling et al. (2003) and in Haase et al. (2001). Hereafter, we refer to these formulations as the MD-FHW model and the MD-HDD model, respectively. Two similar solution approaches, that are adaptations of the approach proposed for the single-depot case in Freling et al. (2003), are also proposed. Both approaches contain two phases: the first phase computes a lower bound on the optimal value, while the second one finds a feasible solution. The lower bound is computed by approximately solving a linear relaxation using a combined column generation/Lagrangian relaxation method. The first approach relies on the linear relaxation of the MD-FHW model while the second one uses that of the MD-HDD model. The second approach also includes a special treatment of the bus-count constraints which are added one at a time. The second phase is identical for both approaches. A heuristic feasible vehicle schedule is found by applying Lagrangian relaxation on the MD-FHW model, where only the duty variables generated during the first phase are considered. Once this schedule is established, a duty schedule is computed for each depot using the DSP approach proposed in Freling et al. (1999, 2003). A series of comparative tests on real-life and randomly generated datasets involving up to 653 segments showed that both integrated approaches can solve these instances to yield substantial savings when compared to the traditional bus-first, duty-second sequential approach. Furthermore, neither of the integrated approaches could clearly outperform the other one, even though the second one regularly provided weaker lower bounds than those produced by the first approach. To reduce solution times and solve larger instances, de Groot and Huisman (2004) devised and

compared different heuristic strategies for splitting an instance into smaller ones which are thereafter solved individually by an integrated approach.

For the multidepot IVDSP, Borndörfer et al. (2004) used an integer programming formulation that essentially combines together model (26)–(30) for the MDVSP and model (32)–(34) for the DSP and adds synchronization constraints between the buses and the drivers on the deadhead, pull-in and pull-out trips. They proposed a heuristic solution approach based on a Lagrangian relaxation of these synchronization constraints. The Lagrangian dual is solved by a proximal bundle method and integer solutions are obtained through a heuristic branch-and-bound procedure. With this approach, they report solving large real-world instances.

4.4 Crew rostering

Given a set of anonymous duties defined over a certain time horizon (typically, a week or a month) for the drivers assigned to a particular depot, crew rostering consists of assigning these duties to the available drivers to form their work schedules (called *rosters*). As with the duties, the validity of the rosters is restricted by safety regulations and collective agreement rules. For instance, a driver cannot work more than a certain number of consecutive days. In most North American public transit agencies, drivers build their own rosters in order of seniority, leaving no place for optimization. On the other hand, in many European agencies, the main objective of the crew rostering problem is to distribute the work load evenly among the drivers, yielding an interesting optimization problem.

As surveyed in Odoni et al. (1994), the common practice for solving transit rostering problems consists of first solving a sequence of assignment problems to build an initial solution and then using a local improvement procedure to better this solution. In the first phase, for each day of the horizon, an assignment problem is defined to assign the duties of the corresponding day to the partial rosters that were built by the previous assignment problems. The cost structure aims at balancing the workload among the drivers. It can also incorporate bonuses to account for the preferences of the drivers for certain duties. An iteration of the second-phase heuristic procedure can be, for example, to select a day, divide all the rosters into two parts according to that day, and solve an assignment problem to match the first parts of the current rosters with possibly different second parts. Such heuristic approaches were developed in the mid-1980s and are still in use due to their computational speed and their flexibility with regards to the work rules.

From a mathematical programming point of view, the crew rostering problem can be formulated as a set partitioning or a set covering problem where a row is defined for each duty and a column is associated with each valid roster. Solving such a model is however not popular for crew rostering problems encountered in public transit systems. This is in contrast to the air and rail contexts where various mathematical programming approaches based on a

set partitioning/covering formulation of the crew rostering problem have been proposed lately in the literature. One major difference between transit and air/rail rostering problems appears to be in the size of the real-world instances, which is larger for transit problems. Indeed, the transit problems are not separable per vehicle type since the drivers are usually allowed to drive all the buses. Furthermore, in these problems, the tasks to cover correspond to individual duties, while they correspond to tours of duties (also known as pairings) that may span up to six days in air/rail rostering problems. Nevertheless, we think that most methodological advances in air/rail crew rostering can be adapted for public transit rostering (at least for small- and medium-size problems). We thus refer the interested reader to Chapters 1 and 2 of this book for a review of the latest advances in air and rail crew rostering.

4.5 Parking and dispatching

An operational planning problem that has attracted little attention in the literature is the management of the parking area in vehicle depots. In congested cities, depots are often restrained in space and quite crowded from late evening to early morning. They also contain different types of buses that are needed by particular bus routes. Therefore, when a bus of a particular type has to leave the depot in the morning, several other buses might need to be moved to clear the way out, resulting in a delay. Two alternatives can be considered to avoid delays. The first one consists of always assigning a directly accessible bus to every morning pull-out even if the bus type is not the one requested for the corresponding bus schedule. In this case, we say that a *mismatch* occurs. The second alternative is to reorder the buses during the night so that they all are properly positioned for the morning pull-outs. An exchange of parking slots between two buses is called a *crossing* or a *maneuver*.

Given a sequence of bus arrivals during the evening, a set of timetabled pull-outs in the morning, and a required bus type for each pull-out, the vehicle parking and dispatching problem consists of parking the buses in the depot upon their arrival and dispatching them to the pull-outs such that the number of mismatches or crossings is minimized while satisfying the following constraints. For safety reasons, buses are not allowed to go backwards in the depot. Therefore, assuming that the depot is made up of lanes that operate as queues, buses enter the lanes at one end and exit them at the other. Obviously, lane capacity must not be exceeded. Finally, given the limited space available to perform crossings, they are only permitted between vehicles of the same lane.

The vehicle parking and dispatching problem is an operational planning problem that usually needs to be solved on a daily basis due to the high variability of the bus availability per type. This variability arises from regular maintenance requirements and unexpected breakdowns. To our knowledge, Winter and Zimmermann (2000) were the first to introduce this problem,

which was defined for tram operations. They showed that it is an NP-hard problem and formulated it as a quadratic assignment model with side constraints. By linearizing this model and adding valid inequalities, they could only solve small-size instances to optimality using the CPLEX MIP solver. Consequently, they proposed heuristics for solving larger instances.

In 2001, Gallo and Di Miele (2001) addressed the vehicle parking and dispatching problem in the context of mass transit buses. They proposed an integer programming model, suitable for both objectives stated above, that relies on three variable types: a first type to assign arriving buses to lanes, a second type to assign morning pull-outs to lanes, and a third type to identify the matchings between the buses and the pull-outs. To solve this model, they developed a three-step heuristic approach. In the first step, Lagrangian decomposition is applied to fix the values of the first two types of variables. In this decomposition, these two types of variables are duplicated to yield two generalized assignment subproblems (one for the arrivals and the other for the pull-outs) that are solved by the CPLEX MIP solver, and a set of *design noncrossing matching* subproblems (one for each lane) which can be solved in polynomial time. In this context, a design noncrossing matching problem consists of matching arriving buses with morning pull-outs with no crossings while selecting, as design decisions, the subsets of buses and pull-outs to consider in the associated lane. A bundle method is used for solving the Lagrangian dual problem. In the second step, after assigning the buses and pull-outs to the lanes according to the last computed solutions in the first step, a simplified design noncrossing matching problem (the design decisions are fixed) is solved for each lane to obtain a complete solution that may contain undesirable crossings or mismatches. A heuristic procedure is then invoked in the third step in an attempt to improve this solution. Using this three-step approach, Gallo and Di Miele (2001) reported solving real-life instances involving up to 4 bus types, 12 lanes, and 77 buses in a few minutes.

Very recently, Hamdouni et al. (2006) argued that an optimal solution to the vehicle parking and dispatching problem, as stated in Winter and Zimmermann (2000) and in Gallo and Di Miele (2001), may be difficult to use in practice due to the randomness of the bus arrival times. Indeed, such a solution may contain a large number of pairs of consecutive slots in a lane to which buses of different types are assigned. Each such pair of slots is likely to lead to a mismatch during the operations if the buses planned for these slots arrive in reverse order. In order to increase the solution robustness, Hamdouni et al. (2006) proposed a restricted definition for the vehicle parking and dispatching problem in which a lane can contain a maximum of two bus types, each bus type being confined to a single block of consecutive parking slots. In this case, a maximum of one pair of consecutive slots per lane is susceptible to mismatches. They also suggest replacing the objective of minimizing the number of crossings by the objective of minimizing the number of lanes that need to be reordered. This suggestion is motivated by the fact that all the buses in a lane must be moved out of the depot when a crossing must be performed

in that lane. Therefore, performing crossings in a lane costs approximately the same independently of the number of crossings to perform. For this version of the problem, Hamdouni et al. (2006) have presented an integer programming model that is based on an enumeration of the possible patterns that can be used to divide a lane into a maximum of two blocks. This model is solved by the CPLEX MIP solver after adding a series of cuts that reduce the feasible region without hindering the search for an optimal solution. Real-world instances involving up to 4 bus types, 16 lanes, and 144 buses were solved to optimality in less than twenty seconds of computation time.

4.6 Maintenance scheduling

Another area where operations research can be helpful in planning public transit operations is bus maintenance scheduling. Since maintenance costs are one of the largest expense categories in a typical transit system, some transit agencies are now investing in maintenance scheduling systems to help them reduce maintenance costs while maintaining a reliable, safe, and attractive transit system. Unfortunately, maintenance systems are not applicable everywhere. When parking depots have limited space, bus assignment is performed on a daily basis according to the day's parking pattern and it is impossible to predict how many miles each bus will travel on the following days. In this case, buses are simply withdrawn from the fleet when they are due for maintenance and maintenance resources are available. These last-minute withdrawals can often reduce service quality. On the other hand, for spacious parking depots where almost all parking slots are directly accessible at any time, it might be desirable to assign buses to specific vehicle schedules that are valid for a long period of time. The rationale behind this strategy is that drivers can then be assigned to the same bus every day in hope that they will be more sensitive to mechanical anomalies of their bus and report minor problems before they become major ones. In this context where vehicle schedules are fixed for a long time horizon, a maintenance system can be devised for scheduling maintenance activities when buses are not supposed to be in service, and in such a way to maximize maintenance resource utilization.

As stated in Haghani and Shafahi (2002), the bus maintenance scheduling problem can be defined as follows. Given the buses' operating schedules, their maintenance requirements, and maintenance resource and crew availability, schedule buses for maintenance and assign them to existing facilities such that each bus is maintained in time, while the amount of time that the buses are out of service is minimized. Maintenance requirements, involving various types of maintenance, are expressed in terms of a maximum mileage or number of days in between two consecutive maintenance activities. Note that maintenance facilities cannot all be used for all types of maintenance.

To our knowledge, the bus maintenance scheduling problem has only been addressed by Haghani and Shafahi (2002). These authors presented three integer programming models for this problem. The first model is very general but

unsolvable in practice. The second one relies on the assumption that the buses requiring maintenance for each type are identified and sorted in the order they should be maintained. This assumption often can be held in practice. The third model is a network model with side constraints that also relies on assumptions that are usually valid in small to medium-size agencies, namely: regular inspections can be performed in all maintenance bays; and, when a bus is due for more than one inspection in a planning period, it is possible to compute for these inspections nonoverlapping maintenance intervals in which these activities will be scheduled. Haghani and Shafahi (2002) proposed three heuristic approaches for solving the second model and a fourth heuristic for the third model. The first two heuristics are simple branch-and-bound methods, while the third heuristic fixes a large number of variables to zero based on the linear relaxation solution, before solving the resulting problem by an exact branch-and-bound scheme. Finally, the fourth heuristic is a network-based algorithm. Using each of these four approaches, Haghani and Shafahi (2002) solved a series of instances arising from simulated operations involving 181 buses and five maintenance types. The results show that all heuristics can solve these instances in less than five minutes on average to yield acceptable solutions. As mentioned by these authors, future research on this problem can focus on developing better heuristic procedures as well as exact solution approaches based on decomposition methods.

5 Real-time control

In actual operations, a wide variety of exogenous and endogenous factors can affect service delivery, such as weather, incidents, variations in traffic conditions, vehicle breakdowns, etc. These factors may degrade the level of service experienced by transit passengers. For this discussion, we differentiate *minor* and *major* service disruptions: minor disruptions are those that create small perturbations from the schedule (e.g., 5–10 minutes), and major disruptions cause longer breaks in the schedule. The distinction is made in order to differentiate the typical responses to these service problems.

To address these challenges, a transit operator may employ a variety of operations control techniques (Turnquist, 1981; Levinson, 1991). These will generally vary depending on the magnitude of the perturbation to service. In normal service with only minor perturbations from the schedule and small service disruptions, vehicle holding and transit signal priority are the most common techniques that are applied. In holding, a vehicle may be held at a stop to improve passenger service. The first part of this section (Section 5.1) addresses the vehicle holding problem. For transit signal priority, transit vehicles may be given preferential treatment in signal timing in order to move more rapidly through a signalized intersection. For reasons of scope, signal priority is not discussed in this chapter.

When major service disruptions occur, more serious control measures may be considered. These can include skipping stops on a route, including *expressing* over parts of the route or skipping particular stops, in order to catch up on the schedule. It may also be useful to re-position a vehicle on the route. In *short-turning*, a vehicle on a route is emptied and placed in service traveling in the other direction on the route, in order to accommodate passenger demand in the other direction. Real-time *deadheading* may also be employed to re-locate a vehicle to another part of the route (or to another route entirely) where it may be of greater service. In addition, extra vehicles and drivers can be made available, to be inserted into service as the need arises. Models and methods for these types of control measures are outlined in Section 5.2.

5.1 Vehicle holding

The transit vehicle holding problem has been explored by many researchers over the past 30 years. Early approaches to this problem have generally focused on either *threshold-based* holding or *schedule-based* holding. The threshold technique involves holding a vehicle only if the preceding headway is below a certain amount of time (e.g., the desired headway or some other threshold value). In this case, the vehicle is held only until the threshold time and then dispatched. If the vehicle arrives after the threshold value, it is dispatched immediately. On the other hand, schedule-based holding involves holding a vehicle only until its scheduled departure time; if it arrives later than the scheduled time, it is dispatched immediately. More recently, in contrast to such holding policies, models have been developed to determine optimal holding times for each vehicle individually. In all of the modeling approaches, the objective is to minimize the total passenger delay (or waiting time), as measured by the delay or waiting time for passengers waiting to board, passengers already on board, and passengers at downstream locations. Generally, for cases where passengers may arrive at stops according to a printed schedule, the delay is measured in terms of the deviation from the schedule. In cases where passengers arrive randomly at the stops, the average passenger waiting time was given by Welding (1957) in the following expression:

$$E[WT] = \frac{E[H]}{2}\left(1 + \frac{\mathrm{Var}[H]}{E[H]^2}\right), \tag{43}$$

where $E[WT]$ is the expected waiting time per person, $E[H]$ is the expected headway, and $\mathrm{Var}[H]$ is the variance of the headway.

Analytic approaches, considering idealized routes with stochastic service characteristics, were studied in the 1970s. The analytic work at this time focused on optimal threshold policies for simplistic networks, as analytic models for optimal threshold-based holding policies were not easily found for more realistic problems. Partly as a consequence, most subsequent analysis schedule-based and threshold-based holding from the 1970s through the 1990s has been conducted using simulation, rather than analytic techniques.

In contrast to methods to find an optimal threshold value, Barnett (1974) introduced a model to solve directly for the optimal holding time of each vehicle at a control stop. Barnett (1974) derived approximations for optimal holding times at a single control point along a transit route, using simplified two-point discrete distributions of the vehicle lateness to that point. The optimal holding time for each vehicle is found by minimizing the expected waiting time, given as a quadratic function of the holding time. The optimal holding times are a function of the mean and variance of the headway distribution, the ratio of the passenger load at the control stop to the load downstream, and the covariance of vehicle arrivals at the control stop.

Turnquist and Blume (1980) extended the analysis by Barnett (1974) to consider the effectiveness of holding decisions. They show that holding will only serve to reduce the passenger waiting time if

$$\mathrm{COV}[H] > \frac{0.5\gamma}{1-\gamma}, \tag{44}$$

where COV[H] is the coefficient of variation of headways at the control stop and γ is the ratio of passengers on board at the control stop to those at downstream stops. The obvious implications of this formula are that control is best implemented when the coefficient of variation of headways is large, and/or when the ratio of on-board passengers to downstream passengers is small. This can be used to select whether and at what locations holding is implemented.

More rigorous analytic methods to solve for individual vehicle holding times have only emerged more recently. Most of the recent holding models include detailed models of transit operations, such as dwell times, passenger boarding and alighting processes, and minimum headway and capacity constraints. For holding decisions, most models assume that vehicle dwell times at stops are modeled explicitly as a linear function of the number of boarding and/or alighting passengers, plus any holding time added at the control stop. More critically, this level of detail allows a certain realism in the modeling of the effect of holding decisions: once a hold is effected, the dwell time of subsequent vehicles at the same stop, and their trajectories downstream, will be changed.

The model formulation of Adamski and Turnau (1998) addressed the problem of minimizing schedule deviations on route. The problem is formulated as an optimal control problem, and the operating dynamics are explained through a set of linear difference equations as a vehicle moves across stops on a route. Using these linear difference equations and a quadratic objective function in the vehicle departure times, the determination of a control (a holding time) can be solved through traditional control methods. This approach appears to be most applicable for maintaining schedule adherence on lower-frequency transit lines, where schedule adherence may be more important than maintaining regular headways.

For higher-frequency service, headway regularity becomes the dominant factor in minimizing passenger waiting time. Holding in this case becomes a problem of adjusting vehicle headways to minimize this variability. Here, we

present the vehicle holding problem formulation based on the work of Eberlein (1995) and Eberlein et al. (2001). A transit route has stations $K = \{1, \ldots, k_t\}$, where k_t is a terminus with sufficient layover time to recover from a minor service disruption. A control is exerted only at stop k, and affects only downstream stops from k to k_t. Let $K' = \{k+1, \ldots, k_t\}$ be the set of downstream stops. Vehicle trips affected by the hold are in the set $I_m = \{i, i+1, \ldots, i+m-1\}$ where trip m is not controlled. The decision variables are the departure times for each vehicle at the control stop ($d_{j,k}$ for vehicle j at the control stop k). The headways, then, are measured as the time between consecutive departures from each stop.

In the model of operations, passengers arrive at the stop k' at a rate of $\lambda_{k'}$. The load on board vehicle j after leaving stop k' is $L_{j,k'}$, and the percentage alighting at any stop k' is given by $q_{k'}$. Also, $a_{j,k'}$ is the arrival time of vehicle j at stop k', $s_{j,k'}$ is the dwell time of vehicle j at stop k' to allow passengers to alight and board, $R_{k'}$ is the running time from stop $k'-1$ to k'. The dwell time is based on a linear function of the alighting and boarding passengers, where c_0 is a constant term, c_1 is the incremental time necessary for one passenger to board, and c_2 is the incremental time for one passenger to alight. Also, delay may propagate at a downstream timepoint k_c if the departure time d_{j,k_c} after passenger alighting and boarding is greater than the scheduled departure time t_{j,k_c}. The departure time d^0_{j,k_c} is the departure time at this timepoint if no control action is taken. Finally, no vehicle may enter a stop for a period of time $h_{\min}$ after a vehicle has departed, and the minimum headway upon entering the stop must be at least h_0.

The formulation is as follows, where θ is a weight on the delay to on-board passengers compared to waiting passengers:

$$\text{minimize} \quad \sum_{j \in I_m} \sum_{k'=k}^{k_t} \frac{\lambda_{k'}}{2} h^2_{j,k'} + \theta L_{i,k}(d_{i,k} - a_{i,k} - s_{i,k}) \tag{45}$$

subject to:

$$d_{j,k} - a_{j,k} - s_{j,k} \geqslant 0, \quad \forall j \in I_m, \tag{46}$$

$$d_{j,k'} - a_{j,k'} - s_{j,k'} = 0, \quad \forall j \in I_m, \forall k' \in K', \tag{47}$$

$$d_{i+m,k} - a_{i+m,k} - s_{i+m,k} = 0, \tag{48}$$

$$a_{j,k'} - d_{j-1,k'} \geqslant h_{\min}, \quad \forall j \in I_m, \forall k' \in K', \tag{49}$$

$$d_{j,k_c} \geqslant \max\{t_{j,k_c}, d^0_{j,k_c}\}, \quad \forall j \in I_m, \tag{50}$$

$$a_{j,k'} = \max\{d_{j,k'-1} + R_{k'}, d_{j-1,k'} + h_0\}, \quad \forall j \in I_m, k' \in K' \cup \{k\}, \tag{51}$$

$$s_{i,k'} = c_0 + c_1 \lambda_{k'} h_{i,k'} + c_2 q_{k'} L_{i,k'-1}, \quad \forall j \in I_m, k' \in K' \cup \{k\}, \tag{52}$$

$$h_{j,k'} = d_{j,k'} - d_{j-1,k'}, \quad \forall j \in I_m, k' \in K' \cup \{k\}, \tag{53}$$

$$L_{j,k'} = \lambda_{k'} h_{j,k'} + (1 - q_{k'}) L_{j,k'-1},$$
$$\forall j \in I_m, k' \in K' \cup \{k\}. \tag{54}$$

The first term in (45) gives the total waiting time experienced by passengers at the current stop and all downstream stops, and the second term is the weighted objective value of delay to passengers on board during the hold. The constraint sets (46)–(54) account for the most commonly used dynamics of operations across the different vehicles and stops. In this formulation, (46)–(48) account for arrival and dwell times at stops; (49)–(50) account for maintaining minimum time separation between vehicles at all stops and at the next timepoint, respectively; (51) accounts for run time between stops ensuring a minimum headway; (52) gives a linear accounting of the dwell time for boarding and alighting passengers at a stop; (53) defines the headways in terms of consecutive departure times; and (54) gives the load on board the vehicle upon leaving a stop.

The formulation in Eberlein et al. (2001) excluded the second term in the objective function (the delay imposed to those on board). The formulation includes as decision variables the departure times $d_{j,k}$ for all vehicles $j \in I_m$ at the control stop k. Note that once these departure times are determined, the system evolution is automatic using (47)–(54). Also, this formulation of the vehicle holding problem is a quadratic program but generally not convex. To solve the problem, a heuristic to optimize each departure time, sequentially across vehicles from i to $i + m$ and iterating until convergence, is proposed.

A similar mathematical formulation is presented by Zhao et al. (2001), but using a more general formulation of the cost function. However, the formulation includes only one decision variable, the holding time of the current vehicle i. The proposed solution technique uses a multiagent system approach, in which the vehicle and the set of impacted stations engage in negotiation using the marginal costs of the proposed holding time. This method is proved to be optimal for convex cost functions.

An extension of these models is described by Sun and Hickman (2004). This work considers the use of multiple holding stations along a route, in order to minimize a weighted sum of passenger waiting costs and on-board delay. A heuristic is proposed which solves for the optimal holding times of each vehicle at its next holding station. The models are solved sequentially, beginning at the holding station furthest downstream and moving upstream along the route. The solution at each holding station is solved using a steepest descent method.

The vehicle holding problem formulation by Hickman (2001) differs from these previous formulations in that the analysis explicitly includes stochastic elements, in contrast to these strictly deterministic models. In Hickman (2001), the passenger arrival and alighting processes and the vehicle running times are considered stochastic. The objective function and the operational dynamics incorporate these processes through the expected values and variances of the vehicle headways and loads. The total passenger waiting time, based on (43),

yields the following objective function, minimizing over the holding time $t \geqslant 0$:

$$\text{minimize} \quad \sum_{k'=k}^{k_t} \frac{\lambda_{k'}}{2} \sum_{j \in I_m} \left(\text{Var}[h_{j,k'}|t] + E[h_{j,k'}|t]^2 \right) + \theta E[L_{i,k}]t, \quad (55)$$

where the variables are defined as before. This objective is minimized, subject to the operational dynamics including both expectations and variances of headway and load. To this end, the operations model of Marguier (1985) was used, in which the expectations and variances of vehicle headways and loads are formulated as linear difference equations. As a result, the model becomes a (convex) quadratic optimization problem with linear constraints, in the single decision variable of the holding time t. A simple line search is proposed to find the optimal holding time, while accounting for these operational dynamics.

Another version of the holding problem considers holding strategies for vehicles at a timed transfer terminal. In this condition, passengers arriving late on one route may not be able to make a transfer to a connecting route. The purpose of terminal holding is to hold a vehicle so that transfer passengers can make a connection. In this situation, the objective includes the delay to passengers on board or downstream on the held route, and the delay to passengers wishing to transfer. A hold reduces the delay to transfer passengers while increasing the delay to passengers on board or downstream. The critical variables include the lateness of vehicles at the terminal and the volume of transfer, boarding, and downstream passengers on each route.

In a deterministic operating environment where the passenger boarding and transfer passenger loads are known, and the lateness of any vehicle is known with certainty, the problem reduces to the situation where either the vehicle is not held, or it is held until the moment when a vehicle on another route arrives. In Hall et al. (2001), this model was extended to stochastic vehicle arrivals, giving a distribution of vehicle lateness on each route. In this case, analytic methods may be used to determine the optimal holding time for each route. The objective function reaches a global minimum either with no holding or at one of the local minima in the neighborhood of each expected vehicle arrival time. An extension of this model to the case of technology that allows one to forecast vehicle arrivals at a transfer terminal is described in Dessouky et al. (2003).

5.2 *Other strategies*

There have been a variety of other control strategies that have been analyzed using mathematical programming techniques. In all the cases cited here, deterministic models of transit operations are used. This means that the objective function uses the waiting time as a function of the square of the headway.

Li et al. (1991, 1992) presented a model in which stop-skipping and holding are considered simultaneously in order to bring a route back on schedule after a service disruption. In this case, skipping stops will reduce dwell times for a

vehicle, reducing the preceding headway. This may reduce the average passenger waiting time, but this effect is weighed against the extra waiting time for passengers whose stops are skipped. With this in mind, the objective function is to minimize the total passenger waiting time along the route, by selecting for each vehicle which stops to include in its trip. The decision variables include binary variables indicating if vehicle j is to stop at stop k, and the continuous variable of the departure time of each vehicle at each stop. Constraints include the vehicle operating dynamics and the vehicle capacity. Three solution heuristics are proposed that iterate among local improvement techniques, with each iteration considering changes in only a subset of decision variables.

Fu et al. (2003) proposed a variation on stop-skipping in which every second vehicle is considered for stop-skipping. In this way, at most one vehicle will pass a stop before it is served. The problem is formulated as a nonlinear 0–1 programming problem, and is solved separately for each dispatched vehicle. The objective function includes passenger waiting time, passenger in-vehicle time, and the bus travel time. Constraints include the typical operations dynamics of passenger boarding and alighting processes and bus running times. The problem is solved using explicit enumeration for each bus. Results suggest that this can be solved in real-time for routes with a small number of potential stops; there were 14 stops in their case study.

These previous studies assume that a stop-skipping decision is made before the vehicle is dispatched from a terminal. Sun and Hickman (2005) extended this concept to consider a real-time stop-skipping policy, made while the vehicle is traveling on the route. Based on the latest vehicle location and disruption information, the problem is formulated as solving for a skipping segment along the route, considering passenger waiting time and in-vehicle time. The model solves for the start and end point of the skipped segment using explicit enumeration. A route with 41 stops was analyzed in their case study, with the model being solved in real time (i.e., in seconds of CPU time).

Eberlein (1995) examined the real-time control actions of expressing and deadheading. The expressing problem is defined as determining if a vehicle should skip over a route segment. In this definition, both the starting stop and ending stop of the *express segment* are determined. However, only one express segment is considered per vehicle. In the deadheading problem, a vehicle is to be dispatched from a terminal, and the decision faced is whether to begin revenue service at the terminal or to begin revenue service further down the route. If deadheading is preferred, a *deadhead segment* is created over which the vehicle runs empty. The deadheading work was later published in a separate paper (Eberlein et al., 1998). Because of similarities in the formulation of these problems, only the deadheading work is presented here.

In Eberlein (1995) and Eberlein et al. (1998), this problem was formulated as a nonlinear integer program in order to identify at what downstream stop to resume revenue service. The complication of vehicle dynamics make an analytic solution impossible and significantly complicates the solution using mathematical programming techniques. However, under simplifying assumptions

about the passenger boarding and alighting processes and the vehicle running times in the deadhead segment, the problem becomes analytically tractable. With this simplification, the objective function (minimizing total passenger waiting time) is convex in the number of stops to skip. Using a continuous relaxation, the end of the deadhead segment is calculated as a real value, which is then rounded up or down to the nearest integer to find the stop that minimizes the objective function.

In Eberlein (1995) and Eberlein et al. (1999), a set of models are proposed to simultaneously examine holding, deadheading, and expressing. It is observed that the strategy of holding has the opposite effect of deadheading and expressing: holding a vehicle will lengthen the preceding headway, while deadheading and expressing a vehicle will shorten the preceding headway. As a result, at most one control strategy will be applied to a single vehicle i. This means the control problems are separable, and an efficient heuristic is applied. In the first step, a holding time is determined, following the heuristic from Eberlein et al. (2001). If station skipping is feasible, deadheading and expressing are considered for vehicle i, and holding is considered for subsequent vehicles $i+1$ to $i+m$ to minimize the total passenger waiting time.

Two other recent studies have examined additional operations control strategies under service disruptions, particularly for rail lines. O'Dell and Wilson (1999) performed a comparison of holding and short-turning. The holding model formulation has as its objective minimizing the passenger waiting time along the route, but uses a piecewise linear function as an approximation of the traditional quadratic objective function. Hard capacity constraints are included, resulting in integer constraints and, as a result, a mixed integer linear program. The short-turning model extends the holding model to accommodate a vehicle that may be turned around on the route at a control stop; it is also formulated as a mixed-integer linear program. Commercial software is used to solve both the holding and short-turning models for a set of disruption scenarios.

Shen and Wilson (2001) extended the model of O'Dell and Wilson (1999) to include expressing, in addition to short-turning and holding. The objective function again includes a piecewise linear approximation to the quadratic function of passenger waiting time, but includes a large number binary variables for whether or not to skip a stop (during expressing) and whether or not to short-turn a vehicle. Additional linear approximations of several nonlinear constraints are used to create a mixed-integer linear program. Commercial mixed-integer programming software is used to solve these models.

Finally, a model has been presented by Li et al. (2004) which considers bus re-routing to accommodate vehicle breakdowns. Such a model can be used to insert a replacement vehicle or re-assign existing service vehicles when a vehicle must unexpectedly be removed from service. The model uses an auction heuristic for the multidepot vehicle scheduling problem (MDVSP) to solve practical instances in real time.

6 Conclusion

This survey chapter has reviewed the operations research literature applied to the domain of public transit, with a focus on recent contributions. It has highlighted a fruitful cooperation between the public transit agencies and the operations research community. Indeed, public transit has provided interesting and challenging problems to operations research, while operations research has been successful at solving efficiently several important public transit problems (for instance, network design, timetabling, vehicle scheduling, and crew scheduling). Research on these problems is still going on with the aim of developing new solution approaches or improving existing ones that will allow to solve larger instances and to address additional complexities such as stochasticity and complicated operational rules that were previously ignored.

This survey has also shown that new problems (integration of vehicle and crew scheduling, bus parking and dispatching, as well as a wide variety of real-time control problems), presenting new challenges to the operations research community, have also been studied recently. Research on these problems has already suggested innovative models and solution methodologies which might be applicable in practice in a near future.

This fruitful collaboration between transit agencies and operations research will certainly continue for a long time as transit agencies continue to strive to provide a good quality service at minimum cost, with continual pressure from budgetary restrictions. Operations research should therefore remain an essential tool for helping the agencies plan and run their operations efficiently.

References

Books from CASPT conferences in chronological order

Wren, A. (Ed.) (1981). *Computer Scheduling of Public Transport*. North-Holland, Amsterdam.

Rousseau, M. (Ed.) (1985). *Computer Scheduling of Public Transport 2*. North-Holland, Amsterdam.

Daduna, J.R., Wren, A. (Eds.) (1988). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 308. Springer-Verlag, Heidelberg.

Desrochers, M., Rousseau, M. (Eds.) (1992). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 386. Springer-Verlag, Heidelberg.

Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.) (1995). *Computer-Aided Transit Scheduling*. *Lectures Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg.

Wilson, N.H.M. (Ed.) (1999). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg.

Voss, S., Daduna, J.R. (Eds.) (2001). *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Heidelberg.

Hickman, M., Mirchandani, P., Voss, S. (Eds.) (in press). *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag, Heidelberg, in press.

Cited references

Adamski, A., Turnau, A. (1998). Simulation support tool for real-time dispatching control in public transport. *Transportation Research – Part A* 32 (2), 73–87.

Andreasson, I. (1977). A method for the analysis of transit networks. In: Roubens, M. (Ed.), *Advances in Operations Research*. North-Holland, pp. 1–8.

Baaj, M.H., Mahmassani, H.S. (1990). TRUST: A LISP program for the analysis of transit route configurations. *Transportation Research Record* 1283, 125–135.

Baaj, M.H., Mahmassani, H.S. (1992). Artificial intelligence-based system representation and search procedures for transit route network design. *Transportation Research Record* 1358, 67–70.

Baaj, M.H., Mahmassani, H.S. (1995). Hybrid route generation heuristic algorithm for the design of transit networks. *Transportation Research – Part C* 3 (1), 31–50.

Barnett, A. (1974). On controlling randomness in transit operations. *Transportation Science* 8, 102–116.

Barnhart, C., Johnson, E.L., Nemhauser, G.L., Savelsbergh, M.W.P., Vance, P.H. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46, 316–329.

Bertossi, A.A., Carraresi, P., Gallo, G. (1987). On some matching problems arising in vehicle scheduling models. *Networks* 17, 271–281.

Bianco, L., Mingozzi, A., Ricciardelli, S. (1994). A set partitioning approach to the multiple depot vehicle scheduling problem. *Optimization Methods and Software* 3, 163–194.

Bielli, M., Caramia, M., Carotenuto, P. (2002). Genetic algorithms in bus network optimization. *Transportation Research – Part C* 10 (1), 19–34.

Bookbinder, J.H., Désilets, A. (1992). Transfer optimization in a transit network. *Transportation Science* 26 (2), 106–118.

Borndörfer, R., Grötschel, M., Löbel, A. (2003). Duty scheduling in public transit. In: Jäger, W., Krebs, H.-J. (Eds.), *Mathematics – Key Technology for the Future*. Springer-Verlag, New York, pp. 653–674.

Borndörfer, R., Löbel, A., Weider, S. (2004). A bundle method for integrated multi-depot vehicle and duty scheduling in public transit. ZIB-Report 04-14, Konrad-Zuse-Zentrum für Informationstchnik, Berlin, Germany. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

Bouzaïene-Ayari, B., Gendreau, M., Nguyen, S. (2001). Modeling bus stops in transit networks: A survey and new formulations. *Transportation Science* 35 (3), 304–321.

Bowman, L.A., Turnquist, M.A. (1981). Service frequency, schedule reliability and passenger wait times at transit stops. *Transportation Research – Part A* 15 (6), 465–471.

Carpaneto, G., Dell'Amico, M., Fischetti, M., Toth, P. (1989). A branch and bound algorithm for the multiple vehicle scheduling problem. *Networks* 19, 531–548.

Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research – Part A* 18 (5–6), 439–453.

Ceder, A. (1986). Methods for creating bus timetables. *Transportation Research – Part A* 21 (1), 59–83.

Ceder, A. (1989). Optimal design of transit short-turn trips. *Transportation Research Record* 1221, 8–22.

Ceder, A., Israeli, Y. (1998). User and operator perspectives in transit network design. *Transportation Research Record* 1623, 3–7.

Ceder, A., Tal, O. (1999). Timetable synchronization for buses. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transport Scheduling*. Springer-Verlag, Heidelberg, pp. 245–258.

Ceder, A., Wilson, N.H.M. (1986). Bus network design. *Transportation Research – Part B* 20 (4), 331–344.

Ceder, A., Golany, B., Tal, O. (2001). Creating bus timetables with maximal synchronization. *Transportation Research – Part A* 35, 913–928.

Chien, S., Schonfeld, P. (1998). Joint optimization of a rail transit line and its feeder bus system. *Journal of Advanced Transportation* 32 (3), 253–284.

Chowdhury, S., Chien, S. (2002). Intermodal transit system coordination. *Transportation Planning and Technology* 25, 257–287.

Chriqui, C., Robillard, P. (1975). Common bus lines. *Transportation Science* 9 (1), 115–121.

Cominetti, R., Correa, J. (2001). Common-lines and passenger assignment in congested transit networks. *Transportation Science* 35 (3), 250–267.

Costa, A., Branco, I., Paixão, J.M.P. (1995). Vehicle scheduling problem with multiple type of vehicles and a single depot. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 115–129.

Curtis, S.D., Smith, B.M., Wren, A. (1999). Forming bus driver schedules using constraint programming. In: *Proceedings of the 1st International Conference on the Practical Application of Constraint Technologies and Logic Programming (PACLP99)*. The Practical Application Company, London, pp. 239–254.

Daduna, J.R., Paixão, J.M.P. (1995). Vehicle scheduling for public mass transit – an overview. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lectures Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 76–90.

Dantzig, G.B., Wolfe, P. (1960). Decomposition principle for linear programs. *Operations Research* 8, 101–111.

de Cea, J., Fernández, E. (1989). Transit assignment to minimal routes: An efficient new algorithm. *Traffic Engineering and Control* 30 (10), 491–494.

de Cea, J., Fernández, E. (1993). Transit assignment for congested public transport systems: An equilibrium model. *Transportation Science* 27 (2), 133–147.

de Cea, J., Fernández, E. (1996). An empirical comparison of equilibrium and non-equilibrium transit assignment models. *Traffic Engineering and Control* 37 (7), 441–445.

de Cea, J., Fernández, E. (2000). Transit-assignment models. In: Hensher, D.A., Button, K.J. (Eds.), *Handbook of Transport Modelling*. *Handbooks in Transport*. Elsevier, Amsterdam, pp. 497–508.

de Groot, S.W., Huisman, D. (2004). Vehicle and crew scheduling: Solving large real-world instances with an integrated approach. Report EI2004-13, Econometric Institute, Erasmus University of Rotterdam, The Netherlands. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

de Silva, A. (2001). Combining constraint programming and linear programming on an example of bus driver scheduling. *Annals of Operations Research* 108, 277–291.

Dell'Amico, M., Fischetti, M., Toth, P. (1993). Heuristic algorithms for the multiple depot vehicle scheduling problem. *Management Science* 39 (1), 115–125.

Desaulniers, G., Desrosiers, J., Ioachim, I., Solomon, M.M., Soumis, F., Villeneuve, D. (1998a). A unified framework for deterministic time constrained vehicle routing and crew scheduling problems. In: Crainic, T.G., Laporte, G. (Eds.), *Fleet Management and Logistics*. Kluwer Academic, Norwell, MA, pp. 57–93.

Desaulniers, G., Lavigne, J., Soumis, F. (1998b). Multi-depot vehicle scheduling with time windows and waiting costs. *European Journal of Operational Research* 111, 479–494.

Desaulniers, G., Desrosiers, J., Solomon, M.M. (2002). Accelerating strategies for column generation methods in vehicle routing and crew scheduling problems. In: Ribeiro, C.C., Hansen, P. (Eds.), *Essays and Surveys in Metaheuristics*. Kluwer Academic, Norwell, MA, pp. 309–324.

Desrochers, M., Soumis, F. (1988). A generalized permanent labeling algorithm for the shortest path problem with time windows. *INFOR* 26, 191–212.

Desrochers, M., Soumis, F. (1989). A column generation approach to the urban transit crew scheduling problem. *Transportation Science* 23, 1–13.

Desrosiers, J., Rousseau, J.-M. (1995). Results obtained with crew-opt: A column generation method for transit crew scheduling. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lectures Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 349–358.

Desrosiers, J., Dumas, Y., Solomon, M.M., Soumis, F. (1995). Time constrained routing and scheduling. In: Ball, M.O., Magnanti, T.L., Monma, C.L., Nemhauser, G.L. (Eds.), *Network Routing*. *Handbooks in Operations Research and Management Science*, vol. 8. Elsevier, Amsterdam, pp. 35–139.

Dessouky, M., Hall, R., Zhang, L., Singh, A. (2003). Real-time control of buses for schedule coordination at a terminal. *Transportation Research – Part A* 37, 145–164.

Dial, R.B. (1967). Transit pathfinder algorithm. *Highway Research Record* 205, 67–85.

du Merle, O., Villeneuve, D., Desrosiers, J., Hansen, P. (1999). Stabilized column generation. *Discrete Mathematics* 194, 229–237.

Dubois, D., Bel, G., Llibre, M. (1979). A set of methods in transportation network synthesis and analysis. *Journal of the Operational Research Society* 30 (9), 797–808.

Eberlein, X.-J. (1995). Real-time control strategies in transit operations: Models and analysis. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Eberlein, X.-J., Wilson, N.H.M., Barnhart, C., Bernstein, D. (1998). The real-time deadheading problem in transit operations control. *Transportation Research – Part B* 32 (2), 77–100.

Eberlein, X.-J., Wilson, N.H.M., Bernstein, D. (1999). Modeling real-time control strategies in public transit operations. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 325–346.

Eberlein, X.-J., Wilson, N.H.M., Bernstein, D. (2001). The holding problem with real-time information available. *Transportation Science* 35 (1), 1–18.

Elhallaoui, I., Villeneuve, D., Soumis, F., Desaulniers, G. (2005). Dynamic aggregation of set partitioning constraints in column generation. *Operations Research* 53 (4), 632–645.

Fan, W., Machemehl, R.B. (2004). Optimal transit route network design problem: Algorithms, implementations, and numerical results. Report SWUTC/04/167244-1, Center for Transportation Research, University of Texas at Austin.

Forbes, M.A., Holt, J.N., Watts, A.M. (1994). An exact algorithm for multiple depot bus scheduling. *European Journal of Operational Research* 72 (1), 115–124.

Fores, S., Proll, L., Wren, A. (2002). TRACS II: A hybrid IP/heuristic driver scheduling system for public transport. *Journal of the Operational Research Society* 53, 1093–1100.

Freling, R., Paixão, J.M.P. (1995). Vehicle scheduling with time constraint. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 130–144.

Freling, R., Wagelmans, A.P.M., Paixão, J.M.P. (1999). An overview of models and techniques for integrating vehicle and crew scheduling. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 441–460.

Freling, R., Huisman, D., Wagelmans, A.P.M. (2001a). Applying an integrated approach to vehicle and crew scheduling in practice. In: Voss, S., Daduna, J.R. (Eds.), *Computer-Aided Scheduling of Public Transport. Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Heidelberg, pp. 73–90.

Freling, R., Wagelmans, A.P.M., Paixão, J.M.P. (2001b). Models and algorithms for single-depot vehicle scheduling. *Transportation Science* 35 (2), 165–180.

Freling, R., Huisman, D., Wagelmans, A.P.M. (2003). Models and algorithms for integration of vehicle and crew scheduling. *Journal of Scheduling* 6, 63–85.

Fu, L., Liu, Q., Calamai, P. (2003). A real-time optimization model for dynamic scheduling of transit operations. In: *The 82nd Annual Meeting of the Transportation Research Board*, Washington, DC, January.

Furth, P.G. (1985). Alternating deadheading in bus route operations. *Transportation Science* 19 (1), 13–28.

Furth, P.G. (1986). Zonal route design for transit corridors. *Transportation Science* 20 (1), 1–12.

Furth, P.G. (1987). Short turning on transit routes. *Transportation Research Record* 1108, 42–52.

Furth, P.G., Wilson, N.H.M. (1982). Setting frequencies on bus routes: Theory and practice. *Transportation Research Record* 818, 1–7.

Gallo, G., Di Miele, F. (2001). Dispatching buses in parking depots. *Transportation Science* 35 (3), 322–330.

Gao, Z., Sun, H., Shan, L.L. (2004). A continuous equilibrium network design model and algorithm for transit systems. *Transportation Research – Part B* 38 (3), 235–250.

Gentile, G., Nguyen, S., Pallottino, S. (2005). Route choice on transit networks with online information at stops. *Transportation Science* 39 (3), 289–297.

Gintner, V., Kliewer, N., Suhl, L. (2004). A crew scheduling approach for public transit enhanced with aspects from vehicle scheduling. DSOR Working Paper WP0407, University of Paderborn, Germany. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.) (2004). *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

Haase, K., Desaulniers, G., Desrosiers, J. (2001). Simultaneous vehicle and crew scheduling in urban mass transit systems. *Transportation Science* 35 (3), 286–303.

Hadjar, A., Marcotte, O., Soumis, F. (2006). A branch-and-cut algorithm for the multiple depot vehicle scheduling problem. *Operations Research* 54 (1), 130–149.

Haghani, A., Shafahi, Y. (2002). Bus maintenance systems and maintenance scheduling: Model formulations and solutions. *Transportation Research – Part A* 36, 453–482.

Hall, R.W. (1985). Vehicle scheduling at a transportation terminal with random delay en route. *Transportation Science* 19 (3), 308–320.

Hall, R.W. (1986). The fastest path through a network with random time-dependent travel times. *Transportation Science* 20 (3), 182–188.

Hall, R., Dessouky, M., Lu, Q. (2001). Optimal holding times at transfer stations. *Computers & Industrial Engineering* 40, 379–397.

Hamdouni, M., Desaulniers, G., Soumis, F., Marcotte, O., Van Putten, M. (2006). Parking and dispatching buses in depots using block patterns. *Transportation Science* 40 (3), 364–377.

Han, A.F., Wilson, N.H.M. (1982). The allocation of buses in heavily utilized networks with overlapping routes. *Transportation Research – Part B* 16 (3), 221–232.

Hasselström, D. (1981). Public transportation planning – a mathematical programming approach. Doctoral dissertation, University of Göteborg, Sweden.

Hickman, M. (2001). An analytic stochastic model for the transit vehicle holding problem. *Transportation Science* 35 (3), 215–237.

Hickman, M.D., Bernstein, D.H. (1997). Transit service and path choice models in stochastic and time-dependent networks. *Transportation Science* 31 (2), 129–146.

Hickman, M.D., Wilson, N.H.M. (1995). Passenger travel time and path choice implications of real-time transit information. *Transportation Research – Part C* 3 (4), 211–226.

Huisman, D., Freling, R., Wagelmans, A.P.M. (2005). Multiple-depot integrated vehicle and crew scheduling. *Transportation Science* 39 (4), 491–502.

Hurdle, V.F. (1973a). Minimum cost schedules for a public transportation route – I. Theory. *Transportation Science* 7 (2), 109–137.

Hurdle, V.F. (1973b). Minimum cost schedules for a public transportation route - II. Examples. *Transportation Science* 7 (2), 138–157.

Irnich, S., Desaulniers, G. (2005). Shortest path problems with resource constraints. In: Desaulniers, G., Desrosiers, J., Solomon, M.M. (Eds.), *Column Generation*. Springer-Verlag, New York, pp. 33–65.

Israeli, Y. (1992). Transit route and scheduling design at the network level. Doctoral dissertation, Technion Israel Institute of Technology, Haifa, Israel.

Israeli, Y., Ceder, A. (1989). Designing transit routes at the network level. In: *Proceedings of the First Vehicle Navigation and Information Systems Conference*. IEEE Vehicular Technology Society, pp. 310–316.

Israeli, Y., Ceder, A. (1995). Transit route design using scheduling and multiobjective programming techniques. In: Daduna, J.R., Branco, I., Piaxão, J. (Eds.), *Computer-Aided Transit Scheduling. Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 56–75.

Israeli, Y., Ceder, A. (1996). Public transportation assignment with passenger strategies for overlapping route choice. In: Lesort, B. (Ed.), *Transportation and Traffic Theory: Proceedings of the 13th International Symposium on Transportation and Traffic Theory*. Pergamon, pp. 561–588.

Jansson, K., Ridderstolpe, B. (1992). A method for the route choice problem in public transport systems. *Transportation Science* 26 (3), 246–251.

Jordan, W.C., Turnquist, M.A. (1979). Zone scheduling of bus routes to improve service reliability. *Transportation Science* 13 (3), 242–268.

Klabjan, D., Johnson, E.L., Nemhauser, G.L., Gelman, E., Ramaswamy, S. (2002). Airline crew scheduling with time windows and plane-count constraints. *Transportation Science* 36, 337–348.

Klemt, W.D., Stemme, W. (1988). Schedule synchronization for public transit networks. In: Daduna, J.R., Wren, A. (Eds.), *Computer-Aided Transit Scheduling*. Springer-Verlag, New York, pp. 327–335.

Kliewer, N., Mellouli, T., Suhl, L. (2006). A time-space network based exact optimization model for multi-depot bus scheduling. *European Journal of Operational Research* 175 (3), 1616–1627.

Knoppers, P., Muller, T. (1995). Optimized transfer opportunities in public transport. *Transportation Science* 29 (1), 101–105.

Koutsopoulos, H.N., Odoni, A., Wilson, N.H.M. (1985). Determination of headways as a function of time varying characteristics on a transit network. In: Rousseau, J.M. (Ed.), *Computer Scheduling of Public Transport 2*. North-Holland, Amsterdam, pp. 391–414.

Kwan, A.S.K., Kwan, R.S.K., Wren, A. (1999). Driver scheduling using genetic algorithms with embedded combinatorial traits. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 81–102.

Kwan, A.S.K., Kwan, R.S.K., Wren, A. (2001). Evolutionary driver scheduling with relief chains. *Evolutionary Computing* 9, 445–460.

Lam, W.H.K., Gao, Z.Y., Chan, K.S., Yang, H. (1999). A stochastic user equilibrium assignment model for congested transit networks. *Transportation Research – Part B* 33, 351–368.

Lam, W.H.K., Zhou, J., Sheng, Z.-H. (2002). A capacity restraint transit assignment with elastic line frequency. *Transportation Research – Part B* 36, 919–938.

Lampkin, W., Saalmans, P.D. (1967). The design of routes, service frequencies, and schedules for a municipal bus undertaking: A case study. *Operational Research Quarterly* 18 (4), 375–397.

Last, A., Leak, S.E. (1976). Transept: A bus model. *Traffic Engineering and Control* 18 (1), 14–20.

le Clercq, F. (1972). A public transport assignment method. *Traffic Engineering and Control* 14 (2), 91–96.

Lee, K.T., Schonfeld, P. (1991). Optimal slack time for timed transfers at a transit terminal. *Journal of Advanced Transportation* 25 (3), 281–308.

Levinson, H. (1991). Supervision strategies for improved reliability of bus routes. In: *Synthesis of Transit Practice 15*, National Cooperative Transit Research and Development Program.

Li, Y., Rousseau, J.-M., Gendreau, M. (1991). Real-time scheduling on a transit bus route: A 0–1 stochastic programming model. Publication 772, Centre de Recherche sur les Transports, Université de Montréal.

Li, Y., Rousseau, J.-M., Wu, F. (1992). Real-time scheduling on a transit bus route. In: Desrochers, M., Rousseau, M. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 386. Springer-Verlag, Heidelberg, pp. 213–235.

Li, J., Mirchandani, P., Borenstein, D. (2004). Parallel auction algorithm for bus rescheduling. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg, in press.

Lo, H., Yip, C.W., Wan, K.H. (2003). Modeling transfers and nonlinear fare structure in multi-modal network. *Transportation Research – Part B* 37 (2), 149–170.

Lo, H., Yip, C.W., Wan, Q.K. (2004). Modeling competitive multi-modal transit services: A nested logit approach. *Transportation Research – Part C* 12, 251–272.

Löbel, A. (1998). Vehicle scheduling in public transit and Lagrangean pricing. *Operations Research* 44 (12), 1637–1649.

Lourenço, H.R., Paixão, J.P., Portugal, R. (2001). Multiobjective metaheuristics for the bus-driver scheduling problem. *Transportation Science* 35 (3), 331–343.

Magnanti, T.L., Wong, R.T. (1984). Network design and transportation planning: Models and algorithms. *Transportation Science* 18 (1), 1–55.

Marguier, P.H.J. (1985). Bus route performance evaluation under stochastic conditions. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Marguier, P.H.J., Ceder, A. (1984). Passenger waiting strategies for overlapping bus routes. *Transportation Science* 18 (3), 207–230.

Mesquita, M., Paixão, J. (1999). Exact algorithms for the multi-depot vehicle scheduling problem based on multicommodity network flow type formulations. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 221–243.

Mingozzi, A., Bianco, L., Ricciardelli, S. (1995). An exact algorithm for combining vehicle trips. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 145–172.

Newell, G.F. (1971). Dispatching policies for a transportation route. *Transportation Science* 5 (1), 91–105.

Nguyen, S., Pallottino, S. (1988). Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research* 37, 176–186.
Nguyen, S., Pallottino, S., Malucelli, F. (2001). A modeling framework for passenger assignment on a transport network with timetables. *Transportation Science* 35 (3), 238–249.
Nielsen, O.A. (2000). A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research – Part B* 34, 377–402.
Nielsen, O.A. (2004). A large scale stochastic multi-class schedule-based transit model with random coefficients. In: Wilson, N.H.M., Nuzzolo, A. (Eds.), *Schedule-Based Dynamic Transit Modeling: Theory and Applications*. Kluwer Academic, Boston, pp. 53–77.
Nuzzulo, A. (2003). Transit path choice and assignment model approaches. In: Lam, W.H.K., Bell, M.G.H. (Eds.), *Advanced Modeling for Transit Operations and Service Planning*. Pergamon, Amsterdam, pp. 93–124.
Nuzzulo, A., Russo, F., Crisalli, U. (2001). A doubly dynamic schedule-based assignment model for transit networks. *Transportation Science* 35 (3), 268–285.
O'Dell, S., Wilson, N.H.M. (1999). Optimal real-time control strategies for rail transit operations during disruptions. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Heidelberg, pp. 299–323.
Odoni, A.R., Rousseau, J.-M., Wilson, N.H.M. (1994). Models in urban and air transportation. In: Pollock, S.M., Rothkopf, M.H., Barnett, A. (Eds.), *Operations Research and the Public Sector*. *Handbooks in Operations Research and Management Science*, vol. 6. North-Holland, Amsterdam, pp. 107–150.
Pattnaik, S.B., Mohan, S., Tom, V.M. (1998). Urban bus transit network design using genetic algorithm. *Journal of Transportation Engineering* 124 (4), 368–375.
Poon, M.H., Wong, S.C., Tong, C.O. (2004). A dynamic schedule-based model for congested transit networks. *Transportation Research – Part B* 38, 343–368.
Rapp, M.H., Gehner, C.D. (1976). Transfer optimization in an interactive graphic system for transit planning. *Transportation Research Record* 619, 27–33.
Ribeiro, C.C., Soumis, F. (1994). A column generation approach to the multiple depot vehicle scheduling problem. *Operations Research* 42 (1), 41–52.
Salzborn, F.J.M. (1972). Optimum bus scheduling. *Transportation Science* 6 (2), 137–148.
Salzborn, F.J.M. (1980). Scheduling bus systems with interchanges. *Transportation Science* 14 (3), 211–220.
Scheele, S. (1980). A supply model for public transit services. *Transportation Research – Part B* 14, 133–146.
Sheffi, Y., Sugiyama, M. (1982). Optimal bus scheduling on a single route. *Transportation Research Record* 895, 46–52.
Shen, S., Wilson, N.H.M. (2001). An optimal integrated real-time disruption control model for rail transit systems. In: Voss, S., Daduna, J. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, vol. 503. Springer-Verlag, Heidelberg, pp. 335–363.
Shen, Y., Kwan, R.S.K. (2001). Tabu search for driver scheduling. In: Voss, S., Daduna, J.R. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Heidelberg, pp. 121–135.
Silman, L.A., Barzily, Z., Passy, U. (1974). Planning the route system for urban buses. *Computers & Operations Research* 1, 210–211.
Site, P.D., Filippi, F. (1998). Service optimization for bus corridors with short-turn strategies and variable vehicle size. *Transportation Research – Part A* 32 (1), 19–38.
Smith, B.M., Wren, A. (1988). A bus crew scheduling system using a set covering formulation. *Transportation Research – Part A* 22, 97–108.
Spiess, H. (1983). On optimal route choice strategies in transit networks. Publication 285, Centre de Recherche sur les Transports, Université de Montréal.
Spiess, H., Florian, M. (1989). Optimal strategies: A new assignment model for transit networks. *Transportation Research – Part B* 23 (2), 83–102.
Stern, H.I., Ceder, A. (1983). An improved lower bound to the minimum fleet size problem. *Transportation Science* 17 (4), 471–477.

Sun, A., Hickman, M. (2004). The holding problem at multiple holding stations. In: Hickman, M., Mirchandani, P., Voss, S. (Eds.), *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economics and Mathematical Systems*. Springer-Verlag, Heidelberg.

Sun, A., Hickman, M. (2005). The real-time stop-skipping problem. *Journal of Intelligent Transportation Systems* 9 (2), 91–109.

Ting, C.-J., Schonfeld, P. (2005). Schedule coordination in a multiple hub transit network. *Journal of Urban Planning and Development* 131 (2), 112–124.

Tom, V.M., Mohan, S. (2003). Transit route network design using frequency coded genetic algorithm. *Journal of Transportation Engineering* 129 (2), 186–195.

Tong, C.O., Richardson, A.J. (1984). A computer model for finding the time-dependent minimum path in a transit system with fixed schedules. *Journal of Advanced Transportation* 18 (2), 145–161.

Tong, C.O., Wong, S.C. (1999). A stochastic transit assignment model using a dynamic schedule-based network. *Transportation Research – Part B* 33, 107–121.

Turnquist, M.A. (1978). A model for investigating the effects of service frequency and reliability on bus passenger waiting times. *Transportation Research Record* 663, 70–73.

Turnquist, M.A. (1981). Strategies for improving reliability of bus service. *Transportation Research Record* 818, 7–13.

Turnquist, M.A., Blume, S.W. (1980). Evaluating potential effectiveness of headway control strategies for transit systems. *Transportation Research Record* 746, 25–29.

van Nes, R., Hamerslag, R., Immers, B.H. (1988). Design of public transport networks. *Transportation Research Record* 1202, 74–83.

Verma, A., Dinghra, S.L. (2005). Feeder bus routes generation within integrated mass transit planning framework. *Journal of Transportation Engineering* 131 (11), 822–834.

Wahba, M., Shalaby, A. (2005). A multi-agent learning-based approach to the transit assignment problem: A prototype. *Transportation Research Record* 1926, 96–105. Paper taken from CD-ROM of Conference Proceedings.

Welding, P.I. (1957). The instability of a close-interval service. *Operational Research Quarterly* 8 (3), 133–148.

Wilson, N.H.M., Nuzzolo, A. (Eds.) (2004). *Schedule-Based Dynamic Transit Modeling: Theory and Applications*. Kluwer Academic, Boston.

Winter, T., Zimmermann, U.T. (2000). Real-time dispatch of trams in storage yards. *Annals of Operations Research* 96, 287–315.

Wirasinghe, S.C., Liu, G. (1995). Optimal schedule design for a transit route with one intermediate time point. *Transportation Planning and Technology* 19, 121–145.

Wren, A., Rousseau, J.-M. (1995). Bus driver scheduling – an overview. In: Daduna, J.R., Branco, I., Paixão, J.M.P. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 430. Springer-Verlag, Heidelberg, pp. 173–187.

Wu, J.H., Florian, M. (1993). A simplicial decomposition method for the transit equilibrium assignment problem. *Annals of Operations Research* 44, 245–260.

Wu, J.H., Florian, M., Marcotte, P. (1994). Transit equilibrium assignment: A model and solution algorithms. *Transportation Science* 28 (3), 193–203.

Zhao, J., Dessouky, M., Bukkapatnam, S. (2001). Distributed holding control of bus transit operations. In: *Proceedings of the IEEE Intelligent Transportation Systems Council (ITSC) Conference*, Oakland, CA, August.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14003-7

Chapter 3

Passenger Railway Optimization

Alberto Caprara
DEIS, Università di Bologna, Viale Risorgimento 2, 40136 Bologna, Italy
E-mail: acaprara@deis.unibo.it

Leo Kroon
NS Reizigers, Department of Logistics, P.O. Box 2025, 3500 HA Utrecht, The Netherlands
and
Erasmus University Rotterdam, Rotterdam School of Management,
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
E-mail: lkroon@rsm.nl

Michele Monaci
DEI, Università di Padova, Via Gradenigo 6/A, 35131 Padova, Italy
E-mail: monaci@dei.unipd.it

Marc Peeters
Electrabel, Strategy R&D, Avenue Einstein 2A, 1348 Louvain-la-Neuve, Belgium
E-mail: marc.peeters2@electrabel.be

Paolo Toth
DEIS, Università di Bologna, Viale Risorgimento 2, 40136 Bologna, Italy
E-mail: ptoth@deis.unibo.it

1 Introduction

Railway systems are highly complex systems. Therefore, planning and operational processes related to railway systems are fields that are rich in interesting combinatorial optimization problems. Well-known examples of these are operational planning problems such as line planning, timetabling, platforming, rolling stock circulation, shunting, and crew planning.

However, in the railway industry it was recognized only recently that such problems can be analyzed and solved through the application of mathematical models and optimization techniques, and that this kind of innovation may lead to an improvement in the performance of the railway system as a whole, as well as to a reduction in the time required for solving these problems. The latter aspect is important, because it increases the flexibility of the railway system: the system can adapt in a faster way to changes in the environment.

Railway transportation can be split into passenger transportation and cargo transportation. In this chapter, we mainly focus on the European situation, where the major part of railway transportation consists of passenger trans-

portation, without addressing important problems in cargo transportation such as car blocking, train makeup, train routing, empty car distribution. Whereas in the recent past many European railway companies were state-owned, many of them are currently operating (partly) on a commercial basis, due to the new regulations of the European Commission, which specify that the management of the infrastructure should be the responsibility of the governments, but operating trains should be carried out by independent Train Operators on a commercial basis. This introduces the separate organizations of the Infrastructure Manager, who is responsible for train planning and real-time traffic control, and the Train Operators, who provide their preferred timetable, rolling stock, and transport services. The Train Operators can be split in operators of passenger trains and operators of cargo trains.

The recognition that combinatorial optimization problems arising in railway applications can be solved through the application of mathematical models and of the corresponding solution techniques was caused by several factors. Indeed, the ability to undertake infrastructure planning in a very timely, smooth and efficient way is becoming one of the most important tasks of the Infrastructure Manager, who at the same time has to optimize the use of the infrastructure and to provide line allocation as well as time slot allocation, through rational and transparent procedures. Moreover, the Train Operators increased the demand for improved performance and for higher speed and flexibility in the planning and in the operations. This also stimulated the search for innovative decision support tools, which were recognized much earlier already in the airline industry. On the other hand, the increased power of the currently available hardware as well as the recent developments in mathematical modeling and solution techniques allow one to find solutions close to optimality for real-world combinatorial optimization problems that, until the end of the last century, one did not dare to dream about. Nevertheless, a lot of research is still to be carried out, since in general the combinatorial optimization problems in the railway industry are highly complex and usually also much larger than the corresponding ones in the airline industry.

As a consequence of the above discussion, real-world applications of mathematical optimization techniques in the railway industry are not yet as widespread as in the airline industry. In this chapter, we focus attention on the optimization techniques developed in each of the operational areas mentioned in the beginning, devoting a section to each of them and considering the problems in the order in which they have to be faced in the planning phase.

We start in Section 2 with the Line Planning Problem, which amounts to deciding the routes for the passenger trains as well as the types and frequencies of the trains on each route. Afterwards, the actual timetable of each train has to be fixed, solving the Train Timetabling Problem of Section 3. Once their timetable is fixed, trains have to be assigned to platforms in the stations they visit, leading to the Train Platforming Problem of Section 4. The definition of train units (locomotives and train carriages) to be assigned to the trains with known timetable and platforms is the Rolling Stock Circulation Problem of

Section 5, whereas the parking of these train units outside the stations when they are not used or are in maintenance is the Train Unit Shunting Problem of Section 6. Finally, the definition of the workload of train drivers and conductors to operate a given timetable is the Crew Planning Problem addressed in Section 7.

Each section contains an introduction, a survey of the available literature, a formal statement of the problem along with one or more (Mixed) Integer Linear Programming ((M)ILP) formulations, and for some cases, experimental results on specific case studies. The (M)ILP formulations presented have been used to solve the associated problems in practice through the solution of the associated Linear Programming (LP) relaxation, either directly by some general-purpose solver or (heuristically) by using Lagrangian relaxation and heuristic methods. Huisman et al. (2005) and Bussieck et al. (1997) also provide overviews of the application of mathematical optimization techniques in passenger railway systems.

Other areas where optimization techniques could be used effectively, such as long term capacity planning of infrastructure, rolling stock, and personnel did not yet get a lot of attention in the literature. The same applies to real-time control of the railway processes: in these areas the development of effective model based decision support tools is still in its infancy.

Nevertheless, we conclude the introduction with a brief description of a number of strategic issues, in order to provide some background information on the relevant problems there. We hope that this will inspire researchers to study these subjects as well, since, in the long run, adequate models and solution techniques for solving these problems may be even more relevant for the quality of the railway system than the effective solution of the operational scheduling problems.

Strategic planning

Strategic issues in railway systems are related to the desirable service level to be provided to the customers (in terms of number of direct connections, frequencies, and reliability), and the capacities of the resources that are required to accomplish these services. The most important resources are the railway infrastructure, the rolling stock, and the crews. Usually enormous amounts of money are involved with the management of these resources.

Strategic issues are far more difficult to handle by optimization approaches than more operational issues due to the extremely long planning horizons that are involved. For example, building a new infrastructure usually takes several years, and once the new infrastructure has been built, it will be operational for several decades. For rolling stock, the situation is similar. These long planning horizons imply a lot of uncertainty, e.g., about the demand for railway transportation in the long run. Therefore, the availability of dependable demand forecasting models is highly important. Models that are to be used for supporting strategic decisions on long term capacities of resources preferably

take into account this uncertainty explicitly, e.g., by applying scenario analysis or stochastic optimization techniques.

In the following, we briefly address the cases of rolling stock and crew management.

Rolling stock management

Railway passenger transportation is characterized by a demand that is usually concentrated in two peak periods per workday. Therefore, when determining the required capacities of the railway system, not only the total demand for railway transportation has to be taken into account, but also the difference between peak demand and off-peak demand. The demand for passenger transportation during the peak periods is usually not symmetric (e.g., in the morning peak period, most passengers travel towards the large cities). This implies that the utilization rate of the rolling stock cannot be as high as one would like.

The required capacity of the rolling stock for passenger transportation is also influenced by the requested service level. Also the difference between first and second class demand is to be taken into account. Another issue in rolling stock management is the maintenance policy to be applied: a more intensive maintenance strategy requires more rolling stock than a less intensive strategy but, on the other hand, will probably lead to more reliable rolling stock, thereby resulting in less disturbances and a higher punctuality in the operations. Although it is clear that such relations exist in practice, it is hard to model them explicitly, since issues like the reliability of railway systems have hardly been quantified so far.

Options that have to be decided upon in rolling stock management are the following: selection of the types of rolling stock to be used, acquisition of new rolling stock, temporarily hiring or leasing of rolling stock, upgrading of existing rolling stock, life time extension of existing rolling stock, selling of redundant rolling stock, destruction of rolling stock that has completed its life cycle, and, as mentioned, the maintenance policy to be applied. For rolling stock to be acquired, the appropriate types and first and second class capacities per unit are relevant issues besides all kinds of technical specifications. Typically, large rolling stock units may be inflexible in the operations and small units may be relatively expensive.

Usually, in strategic issues the objective is to minimize the expected costs over the life cycle of the rolling stock, given the fact that certain criteria with respect to the service for the passengers have to be met. The latter implies that usually a shortage of rolling stock capacity is worse than a surplus, although both are preferably to be avoided. One of the few papers describing the non-operational problem of allocating the available rolling stock capacity among the lines to be operated, based on the passenger demand during the morning rush hours, is Abbink et al. (2004). This chapter also takes into account some robustness aspects, by minimizing the number of different rolling stock types per line.

Crew management

Crew management for train drivers and conductors deals with strategic issues related, e.g., to the locations and the capacities of the crew depots. Also the balance between the capacity for drivers and the capacity for conductors per depot is important, in particular if drivers and conductors are assumed to operate in teams. Moreover, a balance per depot with respect to age, skills, and gender is to be pursued.

The objective of crew management is to establish a matching between the required and the available capacities of the depots. The required capacities of the depots depend on the timetable and the rolling stock circulation, but also on the agreements between management and crews about the structure of the crew workload. The required capacities of the depots can also be influenced by shifting certain amounts of work from one depot to another.

However, the main measures to be taken for creating a matching involve the capacities of the depots. Relevant measures include: hiring additional crews, training crews so that they become more flexible, or moving crews from one depot to another. Financial measures can be taken to stimulate people to retire earlier. In extreme cases, crews can be fired. Agreements about working conditions, workloads, and salaries should be such that the crews are satisfied in their work, since the crews are an important factor for providing services of a high quality.

An important aspect to be taken into account in crew management is the relatively long throughput time of the process of hiring additional crews until they are fully operational. This time varies from about one year for conductors to about two years for drivers. The length of the throughput time is mainly due to the required training for new employees, both theoretical and on-the-job. Obviously, hiring new employees has a long lasting effect on the capacities of the involved depots.

2 Line planning

Many European operators of passenger trains operate a *cyclic* (also called "clockfaced") timetable. That is, a timetable in which the scheduled trains can be partitioned into a number of subsets, where the trains in each subset have the same routes and the same stop stations. The only difference between the trains in each subset are the arrival and departure times: for two consecutive trains in the same subset and for each station in their common set of stop stations, the difference of the departure times of the trains in the same direction is one cycle time (or half a cycle time). A similar statement also holds for the arrival times of the trains.

Each subset of trains in such a partitioning is called a *line*. The trains in a line differ from each other only in their arrival and departure times. The *frequency* of a line denotes the number of trains that run in each direction per cycle on

their common route. A cycle time that one encounters frequently in practice is the *hour*.

Usually, the stop stations of a line are based on the status of the line and on the status of the stations: there are m types of stop stations and m types of lines, and the trains of a line of type τ_l dwell at the stations of type τ_s if and only if $\tau_l \geqslant \tau_s$. For example, there may be Intercity, Interregional, and Regional stations, and the same line types. Then the trains of an Intercity line dwell at the Intercity stations only, and the trains of a Regional line dwell at all stations underway. A line system based on the foregoing assumptions is called a *heterogeneous* line system. Recently, also other line systems have been designed, such as *homogeneous* line systems. However, these will not be discussed in this chapter.

The Line Planning Problem (LPP) is the problem of designing a line system such that all *travel demands* are satisfied and certain objectives are met. There are two main conflicting objectives to be pursued when planning a line system, namely:

(i) maximizing the service towards the passengers, and
(ii) minimizing the operational costs of the railway system.

In the following we define as *direct passengers* the passengers that can travel from their origin station to their destination station without having to change trains. Maximizing the number of direct passengers usually results in long lines: the longer a line, the more direct connections are provided. However, long lines may transfer delays more easily over a wider geographical area, and they may prohibit an efficient allocation of rolling stock: the rolling stock is usually allocated according to the peak demand along the line. Therefore, in a robust and cost-optimal line system, the lines are relatively short, which may force the passengers to transfer from one train to another relatively often. Thus an acceptable trade-off between the two objectives has to be made. Both objectives have been studied, but the combination of the two in one model has not been described yet.

When designing a line system, one has several options for providing sufficient capacity to transport all passengers: lines can be operated with a high frequency and with trains with a relatively low capacity, or with a low frequency with trains with a relatively high capacity. Here also the choice between the allocation of Single Deck or Double Deck rolling stock is relevant. A high frequency on a line also adds to the service towards the passengers.

Finally, it should be noted that in practice there is a bidirectional relation between the travel demand and the operated line system. On the one hand, the operated line system should be such that the full travel demand can be accommodated. On the other hand, a line system that provides a high service towards the passengers may also attract additional passengers, thereby increasing the original travel demand. Thus, the travel demand also depends on the operated line system.

Literature review

There is a limited set of papers dealing with LPP. Most of them deal with one line type only. None of the papers distinguishes between the utilization of Single Deck or Double Deck train units on a line, although this seems to be a relevant choice.

Dienst (1978) describes a branch-and-bound method to find a line system with a maximum number of direct passengers. He assumes that all trains have an infinite capacity, i.e., if there is a line between two nodes, then this line can accommodate all passengers that want to use this line. Thereby the need for a frequency greater than one on a line is eliminated, but it is questionable whether the capacity of the obtained line plan will be sufficient to accommodate all passengers in practice.

Bussieck et al. (1996) also search for a line system with a maximum number of direct travelers. They use decision variables denoting the frequency of each line and assume that all trains have the same fixed capacity. In order to reduce the number of decision variables, they use an aggregation of the decision variables. This aggregation requires the capacity constraints of the trains to be relaxed. The model is solved by first applying several preprocessing techniques, and then by applying a general-purpose ILP solver. The authors describe several valid inequalities to improve the LP lower bound. Bussieck (1998) further extends these methods.

Claessens et al. (1998) study the problem of finding a minimum cost line system. They start with a nonlinear mixed integer model involving binary decision variables for the selection of the lines and additional variables for the frequencies and the train lengths. Since the nonlinearity of the model leads to computational problems, the authors switch to an ILP. The model is solved by a general-purpose ILP solver, after applying several preprocessing techniques.

Goossens et al. (2004) also focus on the design of a minimum cost line system. Their model is similar to the model described by Claessens et al. (1998). In Goossens et al. (2004) a branch-and-cut approach to solve LPP is described. The main ingredients of the algorithm are preprocessing techniques combined with several classes of valid inequalities to improve the LP lower bounds and a number of strategies for variable selection and branching during the branching process. Goossens et al. (2005) consider the problem of designing a line system for several line types simultaneously. In order to reduce the number of decision variables, they combine and disaggregate the origin/destination flows of the passengers. This is justified by the fact that they only consider the objective of cost minimization and that the consequences for the passengers are not taken into account.

Lindner (2000) also studies the minimization of the costs of the line system. He develops a branch-and-bound method for finding a cost-optimal line system. His model also integrates the minimum cost LPP with a model for finding a cyclic timetable. The timetabling part of the model is based on the model

for the Periodic Event Scheduling Problem described by Serafini and Ukovich (1989), see Section 3.2.

Recently, also the objective of minimizing the number of transfers from one train to another has been studied by Scholl (2005). Note that minimizing the number of transfers from one train to another is more complex than maximizing the number of direct passengers, since basically for each origin/destination pair the associated path through the network has to be followed. Scholl (2005) develops a so-called "Switch-and-Ride" network. For real-life instances such networks are huge. Scholl (2005) uses Lagrangian relaxation for obtaining lower bounds, and several heuristic methods for generating feasible solutions.

Model formulation

LPP is a strategic planning problem, since the line system is the basis of the railway services provided by a Train Operator: the variable costs of a Train Operator are determined to a large extent by the design of the line system. The same holds, to a smaller extent, for the Train Operator's income. As a consequence, LPP is a highly complex problem in which many uncertainties have to be dealt with. For modeling LPP many assumptions have to be made, which are described next.

Assumptions

The basic input for LPP consists of the global structure of the railway infrastructure and the expected demand for railway transportation, e.g., per hour. We assume that the forecasted travel demand is represented by a given origin/destination matrix. The bi-directional relationship mentioned earlier may be handled by applying an optimization model iteratively.

Furthermore, in practice the travel demand is usually not stationary nor symmetric: there are peak hours and off-peak hours, where especially during the peak hours the travel demand has a dominating direction. Nevertheless, line systems are usually symmetric in practice. That is, for each pair of stations s_1 and s_2, the number of direct trains from s_1 to s_2 is more or less the same as the number of direct trains from s_2 to s_1. This is due to the fact that the rolling stock is usually operated on a line-by-line basis and circulates along its line. Due to the principle of flow conservation for the rolling stock, such a system does not allow for large deviations from a symmetric line system.

Since usually the peak hours are the bottlenecks in a railway system, Train Operators are mainly interested in designing a line system that can accommodate the travel demand during the peak hours. Therefore, the capacities of the trains should be such that they can accommodate both the passenger flows during the morning peak hours and the passenger flows during the afternoon peak hours. Hence, if $d^{\mathrm{m}}_{s_1,s_2}$ and $d^{\mathrm{m}}_{s_2,s_1}$ denote the passenger flows per hour between stations s_1 and s_2 during the morning peak, and $d^{\mathrm{a}}_{s_1,s_2}$ and $d^{\mathrm{a}}_{s_2,s_1}$ denote the passenger flows per hour between stations s_1 and s_2 during the afternoon

peak, then the lines and the corresponding line capacities should be such that

$$d_{s_1,s_2} = d_{s_2,s_1} = \max\{d^{\mathrm{m}}_{s_1,s_2}, d^{\mathrm{m}}_{s_2,s_1}, d^{\mathrm{a}}_{s_1,s_2}, d^{\mathrm{a}}_{s_2,s_1}\}$$

passengers can be transported between stations s_1 and s_2 during the peak hours. This property should hold for all pairs of stations s_1 and s_2 simultaneously. Note that the models may be adapted to differentiate between the travel demand during the morning peak and during the afternoon peak, and possibly also between the travel demand during the off-peak hours.

Furthermore, it is usually assumed that each passenger uses a pre-specified path through the network. This is not a very strong assumption, since usually each passenger's path through the network is specified by the ticket regulations: each passenger is supposed to travel along the shortest-distance path through the network from her/his origin to her/his destination. This assumption allows one to compute a priori the number of passengers on each edge in the railway network based on the origin/destination matrix.

Next, although it is also possible to deal with different line types simultaneously, see Goossens et al. (2005), the passenger flows are usually assumed to be split per line type. A consequence is that the different line types can be considered apart from each other. For generating such a split of the passenger flows several methods can be used, e.g., the System Split procedure described by Oltrogge (1994) and by Bouma and Oltrogge (1994). This procedure assumes that each passenger switches to a faster train as soon as possible, and switches back to a slower train as late as possible. For example, if a passenger travels from Regional station s_1 to Intercity station s_3, where Intercity station s_2 is the first Intercity station on the shortest path from s_1 to s_3, then this passenger is assumed to travel from station s_1 to station s_2 in one or more Regional trains, and from station s_2 to station s_3 in one or more Intercity trains.

Finally, it is assumed that the lines are simple lines. That is, each line is defined as a simple path in the railway network. In practice, lines may have a more complex structure, for example, due to underway splitting and combining. An advantage of this structure is that splitting and combining of trains increases the number of direct connections without leading to an additional utilization of the infrastructure on the common part of the route. Note that these more complex lines may also be handled by the model described in the following.

The model

We describe a formulation for a single line type, taking into account both the objective of maximizing the number of direct passengers and the objective of minimizing the operational costs.

We assume that the railway network is represented by an undirected graph $G = (V, E)$, where the nodes $v \in V$ represent the stations and the edges $e \in E$ represent the tracks between the stations. Since it is assumed that only a single line type is considered, the trains are assumed to dwell at all stations underway. The set of stations, however, can be subdivided into the set of stations

where a line may start and end, and the set of stations where this is impossible due to a lack of facilities. The travel demand is given by the symmetric origin/destination matrix defined as follows. The set P denotes the unordered pairs of stations with a positive demand. If $p = (p_1, p_2) \in P$ is such a pair, then let d_p be the number of passengers that want to travel between stations p_1 and p_2. Furthermore, for each $p \in P$, let E_p be the set of edges on the shortest path between stations p_1 and p_2, and with some abuse of notation, we write d_e for the total number of passengers that want to travel along edge e, namely $d_e = \sum_{p:e\in E_p} d_p$.

We assume that a set L of *potential lines* is given a priori. The set of edges of line $l \in L$ is denoted by E_l. The objective is to select an appropriate subset from the given set of potential lines. For each line, also a certain frequency and a certain capacity per train are to be selected. This is necessary, because the costs of a line are strongly related to the provided capacity per train. This capacity is a combination of a rolling stock type (Single Deck or Double Deck) and a number of carriages. The different options for the capacity per train are represented by the set C. The set F denotes the set of potential frequencies. Usually, a line can be operated one or two times per cycle time. The capacity of a line equals the capacity per train multiplied by the line's frequency. The number of rolling stock units that are required to operate a line depends on the number of trains on the line and on the lengths of these trains. Furthermore, the number of trains on a line depends on the line's frequency and on its circulation time.

Each feasible combination of a line $l \in L$, a frequency $f \in F$, and a capacity option $c \in C$ is denoted by an index i. The set of indices i is denoted by I. The line index, frequency index and capacity index corresponding to index i are denoted by l_i, f_i, and c_i, respectively. Thus, the capacity of the line corresponding to index i in terms of the number of passengers that can be accommodated by this line equals $f_i c_i$. The operational cost associated with option i is denoted by k_i. These costs are mainly determined by the variable train costs (e.g., the train driver) and by the variable carriage costs (e.g., the conductor(s) and the carriage kilometers).

The main decision variables in the formulation are the binary variables x_i, equal to 1 if and only if line l_i is to be operated with frequency f_i and capacity c_i. Additional decision variables are the variables d_{lp}, representing the number of direct passengers that travel on line l between the pair of stations p. Now the model for LPP reads as follows:

$$\max w_1 \sum_{l\in L} \sum_{p\in P} d_{lp} - w_2 \sum_{i\in I} k_i x_i \tag{1}$$

subject to

$$\sum_{i\in I: l_i = l} x_i \leqslant 1, \quad l \in L, \tag{2}$$

$$\sum_{i \in I: e \in E_{l_i}} f_i c_i x_i \geqslant d_e, \quad e \in E, \tag{3}$$

$$\sum_{p \in P: e \in E_p} d_{lp} \leqslant \sum_{i \in I: l_i = l} f_i c_i x_i, \quad l \in L, e \in E_l, \tag{4}$$

$$\sum_{l \in L: E_p \subset E_l} d_{lp} \leqslant d_p, \quad p \in P, \tag{5}$$

$$x_i \in \{0, 1\}, \quad i \in I, \tag{6}$$

$$d_{lp} \geqslant 0, \quad l \in L, p \in P. \tag{7}$$

The objective function (1) describes the fact that we want to obtain a balance between maximizing the number of direct passengers and minimizing the operational costs. Here w_1 and w_2 are weights that describe the relative importance of the two partial objectives. Constraints (2) specify that for each potential line $l \in L$, at most one appropriate frequency and one appropriate option for the capacity per train are to be selected. Constraints (3) describe that on each track $e \in E$ the provided capacity should be sufficient to accommodate all passengers that travel on this track e. Next, constraints (4) provide a link between the two sets of decision variables: for each line $l \in L$ and for each edge $e \in E_l$, the total number of direct passengers that travel on line l should not exceed the provided capacity $f_i c_i$ on line l. Finally, constraints (5) specify that the total number of direct passengers between the pair of stations p cannot exceed the total travel demand between these stations. Note that the integrality of the variables d_{lp} is not imposed, since this constraint generally does not influence the obtained results.

There are several additional constraints that may be taken into account. For example, one may wish to specify that only fixed numbers of Single Deck and Double Deck rolling stock units are to be allocated to the lines. Another example involves capacity constraints for the tracks or the stations: for each track or station upper bounds on the number of passing trains can be specified. Similarly, for each station an upper bound on the number of starting and ending trains may have to be respected. Such capacity constraints may positively influence the robustness of the railway system. Service towards the passengers may be improved by also specifying for each track a lower bound on the number of passing trains, or by specifying a lower bound on the total number of direct passengers. These additional constraints can be represented in terms of the available decision variables in a straightforward way.

Comments

Note that the line system is designed at a very early stage in the planning process. Therefore, only rough estimates of the travel demand and of the costs can be made. For example, the cost coefficients k_i in the objective function (1) are mainly based on the characteristics of the peak hours, which may give an overestimate of the real costs. This may be overcome to some extent by taking

into account a "known" relationship between the number of carriage kilometers during peak hours and during off-peak hours. Also the required number of train units determined based on the line plan cannot be more than a rough estimate.

With respect to model (1)–(7), several comments are in place. First, note that a disadvantage of the model is its size: the numbers of decision variables and constraints grow very fast with the numbers of origins and destinations and with the number of potential lines. Therefore, all models that have been studied in the literature consider either the objective of maximizing the number of direct passengers, or the objective of minimizing the operational costs. The size of the model also leads to long computation times: most literature reports that only approximate solutions could be found within reasonable amount of computation time.

Second, note that in model (1)–(7) the capacities of the lines may be insufficient to handle all direct passengers on a direct connection. As a consequence, the model has to "select" the direct passengers from the available ones. However, this may lead to an overestimate of the number of direct passengers. For example, suppose there are 100 passengers between stations s_1 and s_2, 100 passengers between s_2 and s_3, and 100 between s_1 and s_3. If the direct lines between stations s_1 and s_3 have a total capacity of 100 passengers, then the model's result will be that there are 200 direct passengers, namely 100 direct passengers between s_1 and s_2 and 100 direct passengers between s_2 and s_3. There are 0 direct passengers from s_1 to s_3, since such passengers would deteriorate the objective. This overestimate of the number of direct passengers is caused by the fact that the passenger behavior is not represented in the model in detail. On the other hand, a detailed description of the passenger behavior would make the model even more complex to solve than it is already in its current form.

In order to overcome this problem to some extent, one may take into account the lengths of the direct connections in the objective function. That is, a long direct connection is more valuable than a short one. Alternatively, one may argue that the capacities of the lines should be such that, if there is at least one direct connection between a certain pair of stations, then all direct connections between these stations should be able to accommodate *all* passengers that want to travel between these stations. This would reflect the fact that passengers prefer to have direct connections, and that Train Operators try to facilitate these through their line capacities as well. In order to model this alternative set of constraints, for each pair of stations $p \in P$, the decision variables d_{lp} are to be replaced by a binary variable y_p equal to 1 if and only if there is a direct connection between the two stations of pair p. Now the model reads as follows:

$$\max w_1 \sum_{p \in P} d_p y_p - w_2 \sum_{i \in I} k_i x_i \tag{8}$$

subject to

$$\sum_{i \in I: e \in E_{l_i}} f_i c_i x_i \geqslant d_e, \quad e \in E, \tag{9}$$

$$y_p \leqslant \sum_{i \in I: E_p \subset E_{l_i}} x_i, \quad p \in P, \tag{10}$$

$$\sum_{i \in I: l_i = l} x_i \leqslant y_p, \quad p \in P, l \in L, E_p \subset E_l, \tag{11}$$

$$\sum_{p' \in P: E_p \subset E_{p'}} d_{p'} y_{p'} \leqslant \sum_{i \in I: E_p \subset E_{l_i}} f_i c_i x_i, \quad p \in P, \tag{12}$$

$$x_i \in \{0, 1\}, \quad i \in I, \tag{13}$$

$$y_p \in \{0, 1\}, \quad p \in P. \tag{14}$$

The objective function (8) describes that the passengers between each pair of stations are either counted completely as direct passengers, or they are not counted at all. Constraints (10) and (11) provide the link between the x-variables and the y-variables. Note that constraints (11) imply constraints (2). Finally, constraints (12) specify that the capacities of the lines should be such that, if there is a direct connection between a certain pair of stations, then all passengers between these stations should be accommodated on a direct line.

3 Train timetabling

The general aim of the Train Timetabling Problem (TTP) is to provide a timetable for a number of trains on a certain part of the railway network. As was indicated already in Section 2, one may distinguish between cyclic timetables and noncyclic timetables.

An advantage of a cyclic railway system is the fact that such a system's timetable is easy to remember for the passengers. For example, at a certain station, the trains heading for a certain direction always leave at x:12 and x:42. On the other hand, a drawback is that such a system is expensive to operate. Even in the periods between the peak hours with low travel demand, more or less the same timetable is operated as during the peak hours. The only way to differentiate the system's capacity between the peak hours and the off-peak hours is to modify the lengths of the trains. The latter impacts both the variable rolling stock costs and the variable crew costs (shorter trains require less conductors).

3.1 Noncyclic timetabling

The noncyclic TTP is especially relevant on heavy-traffic, long-distance corridors, where the capacity of the infrastructure is limited due to greater traffic

densities, and competitive pressure among the Train Operators is expected to increase in the near future. This allows the Infrastructure Manager to allocate "optimally" the train paths requested by all Train Operators and proceed with the overall timetable design process, possibly with final local refinements and minor adjustments, as in the tradition of railway planners. In brief, this allows each Train Operator to submit requests for paths on the given railway line, and allows the Infrastructure Manager to collect all the requests, run the optimization algorithm to allocate (if possible) all of them at maximum profit, and eventually respond to the Train Operators with the proposed plan of the time slot allocation and the relative "access fees".

The essential characteristics of the process can be summarized as follows:

- Each Train Operator has associated with each train a profit (i.e., a priority), an ideal timetable, with an ideal departure time, and tolerances within which it can be changed.
- The optimal allocation is found by maximizing the overall profit for the Infrastructure Manager, i.e., the difference between the profits of the scheduled trains and a cost penalty function, which takes into account the deviations of the final timetables with respect to the ideal ones.

In addition, under the assumption of a competitive market, the process can be iterated if some Train Operator does not accept the solution and asks for a re-evaluation by the Infrastructure Manager, e.g., by using modified path profits.

Literature review

TTP has received considerable attention in the literature. Many references consider MILP formulations in which the arrival and departure times are represented by continuous variables and there are binary variables expressing the order of the train departures from each station.

Szpigel (1973) considers a variant of these models in which the order of the train departures from a station is not represented by binary variables but by disjunctive constraints. Small instances of the problem are solved by branch-and-bound by computing bounds through the relaxation of these disjunctive constraints. Jovanovic and Harker (1991) solve, by branch-and-bound techniques, a version of these models that calls for a feasible schedule rather than for the optimization of a suitable objective function. Cai and Goh (1994) illustrate a constructive greedy heuristic driven by one of these models. Carey and Lockwood (1995) define a heuristic that considers the trains one at a time (in appropriate order), and for each train solves a MILP analogous to these models in order to schedule the train optimally, keeping the path of the previously scheduled trains partially fixed. More precisely, the relative order of the train departures for these trains is kept fixed, whereas their arrival and departure times may be changed. Higgings et al. (1997) define local search, tabu search,

genetic and hybrid heuristics, finding a feasible solution by using a model in the family above.

Brännlund et al. (1998) discretize the time into one-minute *time slots* and subdivide the line into *blocks*. Operational constraints impose that two trains cannot be in the same block in the same time slot. They define an ILP model with a binary variable x_{sbt} each time the timetable constraints allow train t to be in block b in time slot s. This model is not suited for large size instances as those arising for the main European corridors.

Oliveira and Smith (2000) model TTP as a special case of the Job-Shop Scheduling Problem, considering trains as jobs to be scheduled on lines regarded as resources, and present a hybrid algorithm devised under the Constraint Programming paradigm, showing how to adapt this framework in some special real-life applications.

Different ILP models based on a graph representation of the problem were presented by Caprara et al. (2002, 2006). In both papers, time is discretized (i.e., expressed in minutes from 1 to 1440) and Lagrangian relaxation is used to derive bounds on the optimal solution value as well as to drive a heuristic procedure. This approach, whose main features are outlined in the following, produced good relaxations and heuristic solutions also for large-size instances.

Model formulation

In this section we consider the basic version of noncyclic TTP, which considers a single, one-way line linking two major stations, with a number of intermediate stations in between. Let S represent the set of stations, ordered according to the sequence in which they appear along the line for the running direction considered, and T denote the set of trains.

A *timetable* defines, for each train $t \in T$, the departure time from its first station $f_t \in S$, the arrival time at its last station $l_t \in S$, and the arrival and departure times for the intermediate stations $f_t+1, \ldots, l_t-1$. Each train $t \in T$ is assigned on input an *ideal timetable* with departure time d_{ts} for each station $s \in \{f_t, \ldots, l_t - 1\}$ and arrival time a_{ts} for each station $s \in \{f_t + 1, \ldots, l_t\}$, which would be the most desirable timetable for the train, that may however be modified in order to satisfy the line capacity constraints. In particular, one is allowed to slow down each train with respect to its ideal timetable, and/or to increase the stopping time interval at the stations. Moreover, one can modify the departure time of each train from its first station, or even cancel the train. The final solution for the problem will be referred to as the *actual timetable*.

The *line capacity constraints* impose that overtaking between trains occurs only within a station. To this end, a train is allowed to stop in any intermediate station (even if the ideal timetable does not include a stop in that station) to give the possibility to some other train to overtake it. Furthermore, for each station, there are lower bounds on the time interval between two consecutive arrivals and two consecutive departures, respectively. Assuming the speed of a train on a line segment to be constant, this last constraint implicitly imposes

a minimum time interval between two consecutive trains in the line segment connecting two consecutive stations. Note that line capacity constraints, along with the fact that the actual timetable has to be repeated every day, may force some trains to be canceled to obtain a feasible solution.

The objective is to maximize the sum of the profits of the scheduled trains, defined as follows. The *profit* achieved for each train $t \in T$ depends on the train's *ideal profit* π_t, on the *shift* ν_t, defined as the absolute difference between the departure times from the first station in the ideal and actual timetables, and on the *stretch* μ_t, defined as the (nonnegative) difference between the total travel times in the actual and the ideal timetables. According to the charging rules generally adopted by the Infrastructure Manager, the profit for each train t is given by

$$\pi_t - \phi_t(\nu_t) - \gamma_t \mu_t, \tag{15}$$

where $\phi_t(\cdot)$ is a user-defined nondecreasing function penalizing the train shift (with $\phi_t(0) = 0$), and γ_t is a given nonnegative parameter (i.e., the function penalizing the train stretch is assumed to be linear). Typically, the profit function is identical for trains of the same type (e.g., intercity trains) running in the same time interval of the day. If the profit of train t turns out to be nonpositive, it is better not to schedule train t, i.e., to cancel it.

The model

We next outline a mathematical formulation of the problem, calling for a maximum-profit set of paths in a multigraph, as proposed in Caprara et al. (2002). We refer to that paper for a detailed description. Let $G = (V, A)$ be the directed acyclic multigraph in which nodes correspond to arrivals and departures from the stations along the line at some instant (recall that times are discretized) and arcs correspond both to train stops within a station and to train trips from a station to the next one. More specifically, for each station s except the first one, the nodes associated with an arrival at s at some instant are denoted by U^s, and for each station s except the last one, the nodes associated with a departure from s at some instant are denoted by W^s. The arc set is partitioned into arc sets A^t associated with each train $t \in T$. Arcs in A^t from a node $w \in W^{s-1}$ to a node $u \in U^s$ model train t departing from station $s-1$ at the time instant associated with w and arriving at station s at the time instant associated with u. Moreover, arcs in A^t from a node $u \in U^s$ to a node $w \in W^s$ model train t arriving at station s at the time instant associated with u and departing at the time instant associated with w. In addition, there is one artificial source node σ, whose leaving arcs represent train departures from their first station, and one artificial sink node τ, whose entering arcs represent train arrivals at their last station. The definition of the graph G guarantees that every path from σ to τ using arcs in A^t corresponds to a feasible timetable for train t and vice versa.

The objective function can be modeled by associating, for each train t, a (possibly negative) *profit* p_a with each arc $a \in A^t$. The profit associated

with an overall train timetable is given by the sum of the profits of the arcs in the paths of G corresponding to the scheduled trains.

Line capacity constraints impose that certain pairs of arcs, associated with different trains, cannot be selected in the overall solution. In particular, two arcs are called *incompatible* if they either correspond to train arrivals/departures too close in time, or if they correspond to an overtaking between two consecutive stations.

An ILP formulation of TTP is the following. For each $t \in T$ and each arc $a \in A^t$, introduce a binary variable x_a equal to 1 if and only if arc a is selected in an optimal solution, i.e., the path in the solution associated with train t contains arc a. For notational convenience, for each node $v \in V$ and each train $t \in T$, let $\delta_t^+(v)$ and $\delta_t^-(v)$ denote the (possibly empty) sets of arcs in A^t leaving and entering node v, respectively. Finally, let $\mathcal{C}$ denote the (exponentially large) family of maximal subsets C of pairwise incompatible arcs. Then the model reads as follows:

$$\max \sum_{t \in T} \sum_{a \in A^t} p_a x_a \tag{16}$$

subject to

$$\sum_{a \in \delta_t^+(\sigma)} x_a \leqslant 1, \quad t \in T, \tag{17}$$

$$\sum_{a \in \delta_t^-(v)} x_a = \sum_{a \in \delta_t^+(v)} x_a, \quad t \in T, v \in V \setminus \{\sigma, \tau\}, \tag{18}$$

$$\sum_{a \in C} x_a \leqslant 1, \quad C \in \mathcal{C}, \tag{19}$$

$$x_a \in \{0, 1\}, \quad a \in A. \tag{20}$$

The objective function (16) is defined as the sum of the profits of the arcs associated with each path in the solution. Constraints (17) impose that at most one arc associated with a train is selected among those leaving the source node σ, while constraints (18) impose equality on the number of selected arcs associated with a train entering and leaving each arrival or departure node. Consequently, the set of selected arcs associated with a train can either be empty, or define a path from the source to the sink. Finally, clique constraints (19) forbid the simultaneous selection of incompatible arcs, imposing the line capacity constraints.

Note that (16)–(20) is indeed a Multicommodity Flow formulation, in which the commodity index is hidden in the multigraph definition. It can be shown that the associated LP relaxation can be solved efficiently through a cutting plane approach. On the other hand, the very large number of variables for real-life instances of the problem suggests to use an alternative approach. The one used in Caprara et al. (2002) is based on a reformulation in which the line capacity constraints are modeled using variables associated with the nodes

of G, and then relaxed in a Lagrangian way. The main advantage of the alternative formulation is that the Lagrangian relaxation leads to a relaxed problem in which the profits of the x variables are unchanged, whereas Lagrangian penalties are associated with the nodes of G, which are much easier to handle than penalties associated with the arcs. The resulting Lagrangian problem then calls for a set of paths for the trains, each having maximum Lagrangian profit, given by the sum of the original profits for the arcs in the path, minus the sum of the Lagrangian penalties of the nodes visited by the path. If the maximum Lagrangian profit of a path for a train is nonpositive, then the train is not scheduled by the optimal solution of the relaxed problem.

The approach in Caprara et al. (2002) determines near-optimal Lagrangian multipliers through subgradient optimization, and applies a constructive heuristic procedure to determine feasible solutions at each subgradient iteration. In this procedure, trains are scheduled according to decreasing values of the Lagrangian profits of their paths, and each train is assigned a timetable corresponding to the maximum Lagrangian profit path compatible with the previously scheduled trains. The very large number of line capacity constraints that are relaxed in a Lagrangian way is handled according to a so-called *relax-and-cut* framework, explicitly considering each constraint only when it turns out to be violated by the relaxed solution at some iteration of the subgradient procedure.

A discussion of how the model can be modified to handle additional features of real-world applications can be found in Caprara et al. (2006). These features include manual block signaling for managing at most one train on a line segment between two consecutive stations, a maximum number of trains that can be present in a station at the same time, a prescribed timetable for a subset of the trains, and maintenance operations that occupy a line segment for a given period.

Experimental results

We report some computational experiments obtained by applying the basic model described above on a set of instances from FS Rete Ferroviaria Italiana, the Italian Infrastructure Manager. The program was run on a Digital Ultimate Workstation 533 MHz.

The function penalizing the shift in time of train t is defined as $\phi_t(\nu_t) := \alpha_t \nu_t$, i.e., the penalty is linear in the shift. The profit coefficients (π_t, α_t, and γ_t) are identical for the trains of the same type and are reported in Table 1. The main characteristics of the instances are outlined in Table 2, showing the number of trains on input (*# trains*) followed by an array with the number of Eurostar, Euronight, Intercity, Express, Direct, Local, and Freight trains.

Table 3 reports the results obtained with a limit of 1000 subgradient iterations, recording the best heuristic solution value and the best upper bound found before this limit. In the table, we indicate: the sum of the profits achievable by scheduling each train according to its ideal timetable (*ideal prof.*); the

Table 1.
Train profit coefficients depending on the train type

Train type	π_t	α_t	γ_t
Eurostar	200	7	10
Euronight	150	7	10
Intercity	120	6	9
Express	110	5	8
Direct	100	5	8
Local	100	5	6
Freight	100	2	3

Table 2.
Characteristics of the instances

Instance	First stat.	Last stat.	# stat.	# trains
BN-BO	Brennero	Bologna	40	68 (1, 0, 5, 13, 11, 38, 0)
MU-VR	Munich	Verona	48	54 (0, 0, 0, 7, 0, 47, 0)
CH-RM	Chiasso	Rome	73	36 (15, 0, 9, 0, 0, 7, 5)
MO-MI-1	Modane	Milan	39	16 (1, 0, 3, 0, 4, 5, 3)
MO-MI-2	Modane	Milan	39	23 (0, 3, 1, 11, 0, 2, 6)
PC-BO-1	Piacenza	Bologna	16	221 (28, 3, 52, 35, 28, 14, 61)
PC-BO-2	Piacenza	Bologna	16	93 (12, 3, 18, 10, 20, 6, 24)
PC-BO-3	Piacenza	Bologna	16	60 (12, 1, 14, 0, 12, 7, 14)
PC-BO-4	Piacenza	Bologna	16	40 (6, 0, 10, 0, 12, 2, 10)
BZ-VR	Bolzano	Verona	20	128 (34, 0, 1, 10, 11, 38, 34)
CH-MI	Chiasso	Milan	15	194 (20, 1, 29, 8, 19, 66, 51)

best upper bound found by the subgradient optimization procedure (*best UB*) with, in brackets, the percentage improvement over *ideal prof.*; the solution value found by scheduling the trains by decreasing values of π_j (breaking ties arbitrarily) and assigning to each train the timetable corresponding to the maximum profit path compatible with the previous trains (*greedy sol.*); the value of the best solution found by the Lagrangian heuristic (*best sol.*) with, in brackets, the percentage improvement over *greedy sol.*; the percentage gap between the value of the best heuristic solution and the best upper bound (*GAP%*); the number of trains scheduled in the best solution (*# sched*); the average shift and the average stretch (in minutes) for the trains scheduled in the best solution (*avg.* ν and *avg.* μ); the overall running time in seconds needed to find the best solution (*time*).

The manual methods proceed in a way similar to the one used to derive the solution whose value is reported in the *greedy sol.* column, therefore the quality of the solution provided by the practitioners is typically close to the value given in this entry.

Table 3.
Results on real-world instances

Instance	Ideal prof.	Best UB	Greedy sol.	Best sol.	GAP%	# sched.	Avg. ν	Avg. μ	Time
BN-BO	7130	6891 (3.4%)	6746	6779 (0.5%)	1.7%	68	0.6	0.3	34
MU-VR	5470	4991 (8.8%)	3332	4208 (20.8%)	18.6%	48	1.4	1.1	92
CH-RM	5280	5129 (2.9%)	4844	4871 (0.6%)	5.3%	35	1.9	1.3	48
MO-MI-1	1760	1713 (2.7%)	1648	1684 (2.1%)	1.7%	16	0.6	0.4	0
MO-MI-2	2580	2533 (1.8%)	2486	2520 (1.3%)	0.6%	23	0.9	0.3	78
PC-BO-1	25,740	24,142 (6.2%)	19,535	21,397 (8.7%)	12.8%	192	0.8	0.8	672
PC-BO-2	11,010	10,947 (0.6%)	10,848	10,882 (0.3%)	0.6%	93	0.0	0.2	118
PC-BO-3	7450	7222 (3.1%)	6709	7119 (5.8%)	1.4%	60	0.6	0.8	271
PC-BO-4	4800	4130 (14.0%)	3350	3656 (8.4%)	13.0%	35	3.2	0.9	266
BZ-VR	16,300	16,101 (1.2%)	15,902	16,003 (0.6%)	0.6%	127	0.3	0.0	266
CH-MI	21,930	21,467 (2.1%)	20,923	21,215 (1.4%)	1.2%	193	0.7	0.1	285

The method succeeded in scheduling almost all trains for most instances, the exceptions being instances MU-VR, PC-BO-1, and PC-BO-4. Moreover, for most instances, the final percentage gap is quite small (less than 2%). For the remaining cases, the best heuristic solution is considerably better than the greedy one (with the only exception of instance CH-RM).

3.2 Cyclic timetabling

In a cyclic timetable, each trip is operated in a cyclic way. That is, each period of the timetable is the same. If the cycle time of the timetable is denoted by $\mathcal{T}$, then this means that, if a trip between stations s_1 and s_2 leaves at time t_1 and arrives at time t_2, then analogous trips are carried out with departure and arrival times $t_1 + k\mathcal{T}$ and $t_2 + k\mathcal{T}$ for all integer values of k.

The first ones to develop a model for generating cyclic timetables were Serafini and Ukovich (1989); in that paper, a mathematical model for the so-called Periodic Event Scheduling Problem (PESP) is presented. In PESP, a set of repetitive events is scheduled under cyclic time window constraints. Consequently, the events are scheduled for one cycle in such a way that the cycle can be repeated. Most models for cyclic TTP are based on PESP.

A timetable consists of a number of processes, such as running between two stations, dwelling at a station, passenger connections etc. The start and end times of these processes are the events of the timetable. The set of events is denoted by E. In PESP, for each $e \in E$, the decision variable v_e represents the time instant at which event e has to be scheduled. All constraints that have to be satisfied by these decision variables specify minimum and maximum process times for the corresponding processes between the events. As a consequence, all constraints have the following form:

$$l_{ef} \leqslant (v_e - v_f) \bmod \mathcal{T} \leqslant u_{ef}. \tag{21}$$

Here e and f are two events in the timetable, and l_{ef} and u_{ef} are appropriate lower and upper bounds for the process time of the process between the events e and f. For example, if e and f are the departure and arrival of a train at two consecutive stations, then l_{ef} and u_{ef} denote the minimum and maximum running time of the train on this trip, respectively. The modulo operator models the cyclicity of the timetable. For instance, if the cycle time equals 60 minutes, then the process time from $t = 55$ to $t = 5$ equals 10 minutes.

Since the modulo operation in (21) is relatively hard to handle in optimization methods, it can be replaced by introducing a binary variable q_{ef} for each constraint of the form (21), and by replacing (21) by the following constraint:

$$l_{ef} \leqslant v_e - v_f + \mathcal{T} q_{ef} \leqslant u_{ef}. \tag{22}$$

Variables q_{ef} make the PESP quite hard to solve by standard branch-and-bound methods: due to the relatively large coefficient $\mathcal{T}$ in (22), the LP relaxations of models based on this formulation of PESP are quite weak.

Therefore Schrijver and Steenbeek (1994) develop a constraint propagation algorithm for solving PESP. Their algorithm has been implemented in the DONS system, that has become an indispensable tool in the Dutch long term railway timetabling process of Netherlands Railways and ProRail. Schrijver and Steenbeek also develop local optimization techniques to improve a feasible solution for fixed values of the variables q_{ef} in (22). Instances with up to 250 trains (all trains running in one hour of the Dutch timetable) can be solved usually within reasonable computing time.

In order to cope with the weak LP relaxation of models based on constraints (22), Nachtigall (1999), Lindner (2000), and Peeters (2003) also describe a formulation of PESP based on cycle bases. This formulation is somewhat easier to solve than the formulation based on constraints (22), because of the lower number of integer variables and the somewhat better LP relaxation.

Nachtigall and Voget (1996) use PESP to generate cyclic timetables with minimal passenger waiting times. Odijk (1996) uses the PESP at a strategic level to determine the capacity of the infrastructure around railway stations. Kroon and Peeters (2003) describe a PESP model including variable trip times. This may result in additional flexibility leading to a higher probability of obtaining a feasible solution.

Finally, Kroon et al. (2005) describe a stochastic optimization variant of PESP. Their model explicitly takes into account stochastic disturbances of the railway processes, distinguishing between a planned timetable and several realizations of the timetable under pre-determined stochastic disturbances. The model can be used to allocate time supplements and buffer times to the processes in the planned timetable in such a way that the average delay of the realizations of the trains is minimal. In order to keep the computation times at an acceptable level, they start with an existing timetable and they fix the variables q_{ef} in (22). They show that, by taking into account stochastic disturbances in the design of the timetable, an increase in robustness can be achieved.

4 Train platforming

The definition of the optimal timetables discussed in Section 3 does not take into account the actual routing of the trains within the stations considered. As briefly mentioned in that section, only an upper bound on the maximum number of trains that can be simultaneously present in a station is imposed. Routing a train in a railway station means finding for each train a path from the point where it enters the station to the point where it leaves the station. Thereby a train usually passes through (and possibly stops at) a *platform* within the station. For this reason, the problem is generally referred to as the Train Platforming Problem (TPP).

While the problem is very easy to solve for relatively small stations, in which there is a very small number of alternative paths to route the trains, it becomes challenging for major stations, that typically have very complex topologies.

The customary problem input for a given station contains a set of *directions* for train arrivals and departures, and a set of *platforms* for train stops. There may also be dummy platforms representing the possibility to traverse the station without stopping at any real platform. Moreover, there is a set of trains, each associated with scheduled (or *ideal*) arrival and departure times, arrival and departure directions, and a set of platforms to which the trains may be assigned (possibly without stopping). If a train comes from (or goes to) the shunting area, there is generally no scheduled arrival time, but there may be a maximum time by which the train should arrive at (or depart from) its platform. Finally, the (complex) station topology is generally represented by defining a set of *(bidirectional) routes* joining each direction to the set of platforms that can be reached by that direction. The physical overlap between two routes, that may correspond to the routes sharing either a line segment or a node (or crossing point) is indicated by an *incompatibility* relation between the two routes. It is assumed that each train occupies its associated route for a given time interval, corresponding to the time required by the train to go from its direction to its platform, or vice versa.

There are two types of operational constraints. The first imposes that, for each platform, a minimum time interval must elapse between the departure of a train from the platform and the arrival of the next train at the platform. The second forbids the occupation of an incompatible pair of routes by two trains in the same time instant. In some cases, for major stations with great traffic densities, this latter type of constraints is relaxed, allowing two incompatible routes to be used for an interval whose length does not exceed a given upper bound, at the cost of paying a suitable penalty. However, in the version of the problem discussed here we will not consider this possible relaxation.

The problem requires the specification, for each train, of the platform to which it is assigned along with the associated arrival and departure routes and the actual arrival and departure times that may be different from the ideal ones at the cost of paying a penalty. Generally, using a terminology analogous to that used in Section 3, a maximum (positive or negative) *shift* with respect to these ideal times is allowed. Moreover, note that the choice of the arrival and departure routes implicitly defines the platform to which the train is assigned. As different platforms within those feasible for the train may have different priorities, a penalty is paid if the train is not assigned to its highest-priority platform.

Literature review

TPP has not received considerable attention in the literature so far. De Luca Cardillo and Mione (1998) consider the simplified version in which, for each train, the scheduled arrival and departure times cannot be changed, and the arrival and departure routes are uniquely determined by the choice of the platform. In this case, one may avoid considering the routes explicitly, implicitly

representing their incompatibilities by defining a list of incompatible train-platform pairs of the form (j, a, k, b), stating that it is infeasible to assign train j to platform a as well as train k to platform b. This version of the problem is modeled as a Graph Coloring Problem with additional constraints, for which a heuristic algorithm is proposed and applied to real-world instances. The same version of TPP is addressed by Billionnet (2003), who considers a classical ILP formulation for Graph Coloring and shows how to incorporate into the model the clique constraints associated with the list of incompatible train-platform pairs. Randomly-generated instances are solved by using a general-purpose ILP solver.

A more general version of the problem, in which arrival and departure times and arrival and departure routes are not fixed a priori is addressed in Zwaneveld (1997), Zwaneveld et al. (1996), Zwaneveld et al. (2001), and Kroon et al. (1997). Actually, these authors distinguish among *inbound*, *platform*, and *outbound* routes, meaning the path followed by a train to enter the station (without going to a platform), the path inside the station stopping at a platform, and the path to exit the station, respectively. Although in practice each train has to be assigned all these three paths, the authors consider the possibility that only some of these (possibly none) are found for the train. This version is modeled as a Stable Set Problem on a graph in which each node corresponds to a choice for a train, with an associated cost, and edges join node pairs associated with the same train as well as node pairs corresponding to incompatible choices for different trains. This model is analogous to that presented in the next section. The problem is naturally formulated as an ILP with one binary variable for each node and clique inequality constraints. The above mentioned references describe heuristic and exact (branch-and-cut) algorithms based on the solution of the corresponding LP relaxation. The approach is embedded within the STATIONS system used by the Dutch railway company to solve the problem. Results for real-world instances corresponding to the main Dutch stations are reported in Zwaneveld (1997).

The version addressed in Carey and Carville (2003) is intermediate between the two versions above, in that the arrivals and departure times can be changed but the assignment of a train to a platform uniquely determines the routes that the train will follow on its way to and from the platform. The paper discusses in great detail the conflicts that may arise from the assignments and the procedures that are followed to evaluate the costs. Rather than a mathematical formulation, the authors describe a constructive heuristic that is applied to a real-world instance from the British railway company concerning the station of Leeds.

Model formulation

As anticipated, we describe a version of TPP analogous to the one considered by Zwaneveld et al. (1996). As for TTP in Section 3, times are discretized and expressed in minutes.

Let P denote the set of platforms in the station, D the set of possible directions for train arrivals and departures, and R the set of all (bidirectional) routes. For each direction $d \in D$ and platform $p \in P$, let $R_{dp} \subseteq R$ denote the set of routes linking direction d to platform p and vice versa. For each route $r \in R$, a list $I_r \subseteq R$ of *incompatible* routes is given, specifying the routes that cannot be occupied by a train when r is occupied.

Let T denote the set of trains which have to be assigned to a platform in the time horizon considered. Each train $t \in T$ is associated with an arrival direction $d_t^{\mathrm{a}} \in D$, a departure direction $d_t^{\mathrm{d}} \in D$, and a set $C_t \subseteq P$ of candidate platforms; for each candidate platform $p \in C_t$ a nonnegative (possibly null) penalty c_{pt} for the assignment of train t to platform p is given. If train t is assigned a platform $p \notin C_t$, the penalty is c_{0t}.

In addition, each train $t \in T$ has an associated ideal arrival time u_t^{a} at a platform, along with a maximum arrival shift s_t^{a}, and an associated ideal departure time u_t^{d} from the platform, along with a maximum departure shift s_t^{d}. As the values of these shifts are typically small, often any combination of actual arrival and departure times that do not differ from the ideal ones by more than the corresponding maximum shift is feasible (also allowing, e.g., an arrival s_t^{a} minutes after the ideal time and a departure s_t^{d} minutes before the ideal time). On the other hand, the mathematical model presented in the following can be adapted to handle any constraint on the combination of actual arrival and departure times. Penalties c_t^{a} and c_t^{d} for each minute of shift in the arrival and departure time, respectively, are incurred. As already mentioned, there are trains that arrive from the depot or depart for the depot, for which the values c_t^{a} or c_t^{d} are "small" (or null) and the values of s_t^{a} and s_t^{d} are "large".

The problem requires to define, for each train $t \in T$, a *path*, which is specified by a platform $p \in C_t$, an arrival route $r_1 \in R_{d_t^{\mathrm{a}},p}$, a departure route $r_2 \in R_{d_t^{\mathrm{d}},p}$ and the corresponding arrival and departure times, within s_t^{a} and s_t^{d} from the ideal ones, respectively, so as to minimize the corresponding penalties. Let $\mathcal{P}_t$ denote the set of all feasible paths for train t and, for each path $P \in \mathcal{P}_t$, q_{tP} be the associated penalty, computed according to the above rules. Even for large size instances, the number of possible paths for a train is relatively small, and all these paths can be generated and considered explicitly, due to the relatively small number of routes joining directions to platforms and the limited maximum shifts from the ideal arrival and departure times, that are associated with a timetable determined in previous phases. According to the operational constraints, two paths are incompatible if they either occupy the same platform for time periods that overlap or are too close in time, or use incompatible routes for the same time instant, with the convention that each arrival route $r \in R$ is occupied by a train $t \in T$ for a time w_{rt}^{a} ending at the arrival of the train at the platform, and each departure route $r \in R$ is occupied by a train $t \in T$ for a time w_{rt}^{d} starting from the departure of the train from the platform.

As customary, the incompatibilities between train paths can be represented by an *incompatibility graph* $G = (V, E)$ having one node (t, P) for each train $t \in T$ and path $P \in \mathcal{P}_t$ and edges joining nodes associated either with the same train or with paths that are incompatible. Clearly, each feasible solution of the problem corresponds to a stable set of G that contains one node associated with each train in T, and vice versa. Let $\mathcal{C}$ denote the (exponentially large) family of maximal cliques C of G.

The natural ILP model is obtained by introducing for each node (t, P) of G a binary variable x_{tP} whose value is 1 if and only if train t is assigned path P. The model reads as follows:

$$\min \sum_{t \in T} \sum_{P \in P_t} q_{tP} x_{tP} \tag{23}$$

subject to

$$\sum_{P \in \mathcal{P}_t} x_{tP} = 1, \quad t \in T, \tag{24}$$

$$\sum_{(t,P) \in C} x_{tP} \leqslant 1, \quad C \in \mathcal{C}, \tag{25}$$

$$x_{tP} \in \{0, 1\}, \quad t \in T, P \in \mathcal{P}_t. \tag{26}$$

The objective function (23) simply minimizes the overall cost. Constraints (24) ensure that each train is assigned a path, while constraints (25) forbid the selection of incompatible paths.

The model considered by Zwaneveld et al. (1996) and Zwaneveld et al. (2001) is analogous to the one above, with the difference that a train path is split into three parts as discussed above and a train may not be assigned all these three parts (very loosely speaking, constraint (24) is an inequality). The above references illustrate preprocessing rules that successfully reduce the number of variables and a procedure to generate a relevant subset of polynomially-many clique inequalities (25). All studied instances related to several stations in the Netherlands with up to 50 trains per hour can be solved within small computing time.

5 Rolling stock circulation

The Rolling Stock Circulation Problem (RSCP) is an important problem for Train Operators, since the acquisition of rolling stock is expensive and a long-term investment. Also the operational costs of rolling stock are usually substantial. These costs include maintenance costs and power supply (electricity or diesel). Both are positively correlated to the number of kilometers that the rolling stock travels in the circulation. For these reasons, a Train Operator has to decide carefully on the type and the number of rolling stock units per scheduled train. Other important concerns in the planning of the rolling stock

circulation are the provided service to the passengers, which can give rise to higher revenues, and the robustness of the circulation.

In order to obtain a better match between the available rolling stock and the passengers' seat demand, the compositions of the trains usually can be changed at several stations by adding equipment to or removing it from the trains. These coupling and uncoupling operations are usually penalized with switching costs. The removed equipment can later be used for another train departing from the same station. Several restrictions must be taken into account when changing the composition of a train. These restrictions are related to the time required to carry out the shunting operations to change the composition and the available time at the station, which is the waiting time between two consecutive trips of the train. For this reason, the order of the equipment in the train composition matters, since switching equipment situated in the body of the train requires more time than switching equipment situated at the tail or the head.

Various versions of RSCP arise depending on the equipment that is used and on the nature of the railway network. Concerning the equipment, we distinguish two cases:

(i) locomotives and train carriages, and
(ii) aggregated modules, subsequently called *train units*.

The latter are composed of a number carriages in a fixed composition, and can move in both directions, without the need of an extra locomotive. An example of the latter is the well known French Train à Grande Vitesse (TGV). A TGV train unit consists of six to ten passenger carriages and two power units, each including a driver's cabin, situated at both ends of the train unit, see Ben-Khedher et al. (1998). Train units can move individually in both directions. A scheduled train can then be composed of several coupled train units.

In the first case, for each trip scheduled during the time horizon, one must determine, for the associated train, the locomotive types and their number, and the carriage types and their number. These numbers are not independent of each other, since the number of carriages determines the type and the number of locomotives in order to provide sufficient pulling capacity. Turning a locomotive-hauled train at the endpoint of a line is quite complex, since the locomotive(s) must be uncoupled and driven to the other side of the train, where it must be coupled again. It also occurs that a reserve locomotive is available for the return trip, or that the entire train is turned around, if the so-called Y-shaped tracks are available, see Lingaya et al. (2002). In the second case (train units), one has to determine only the type and the number of train units per type that are deployed. The order of the train units in a train may also be important. Since a train unit can move in both directions, quick turn-around times at the end points are possible.

Concerning the nature of the network, we distinguish two cases as well:

(i) a sparse network with long distances, and hence, long travel times and relatively low frequencies of trains, and

(ii) a dense network with relatively short-distances and high frequencies of trains.

This difference is important, in particular in connection with the preventive maintenance of the rolling stock.

In a sparse network the rolling stock circulation usually describes in detail the circulation of the individual rolling stock units over a longer period of time, thereby taking into account the fact that each individual rolling stock unit should reach a maintenance center sufficiently often. In contrast, in a dense network the rolling stock circulation is usually anonymous. Moreover, there is a certain capacity reservation for preventive maintenance of the rolling stock. Routing the units that need preventive maintenance to a maintenance center is handled in the operations by exchanging duties of rolling stock units. This is possible because, in a dense network, there are usually sufficient possibilities for exchanging units, i.e., time intervals where two identical rolling stock units are located at the same station.

A second difference between dense and sparse networks is that in a dense network with a lot of commuter trains, a seat reservation system usually does not exist and only the expected numbers of passengers are known. On the other hand, in a sparse network, a seat reservation system often exists, such that the Train Operator has detailed knowledge of the number of passengers and can change its circulation based on the actual reservations. Also revenue management becomes possible in such a network.

5.1 Dense network

Literature review

For the dense network case, Schrijver (1993) describes a model to determine the minimum number of train units that must be deployed on a single line in order to avoid seat shortages. A line is defined by two endpoints between which several trains run up and down according to the timetable. In the case considered, the railway company runs an hourly service and one type of train units is deployed there. There exist, however, two subtypes of train units that differ in length and in capacity of first and second class seats. A train can consist of several units of these subtypes. For every trip, the required numbers of first and second class seats are known. Train units can be coupled to or uncoupled from a train at several stations along the line. Obviously, a train unit can only be coupled to a train at a certain station if the unit is available there at the right moment. The model is basically an Integer Multicommodity Flow model with several additional constraints.

The model proposed by Schrijver (1993) assumes that a train composition can change to any other composition at a station between two subsequent trips. In practice, however, several (un)coupling constraints must be taken into account. Therefore, not only the number of units deployed on a trip is important, but also their order in the train composition. In this particular case, only one

operation is allowed, i.e., coupling or uncoupling, but not both. In addition, the position where train units are coupled and uncoupled is fixed, i.e., in front or at the rear of the train, which depends on the station. Moreover, Schrijver (1993) assumes that the allocated rolling stock capacities per train should be such that all passengers should obtain a seat. However, in the absence of a reservation system, the passengers' demand has a stochastic nature. Therefore, a seat for all passengers cannot be guaranteed. A final issue not dealt with by Schrijver (1993) is that, if a train arrives at its endpoint, it leaves the station usually as soon as possible. That is, the train carries out the first trip of the same line leaving the station, possibly after coupling or uncoupling some units. Given that the subsequent trip for a train at an endpoint is known, several sequences of trips can be distinguished. We refer to such a sequence of trips as a *train*.

In order to cope with the above issues, Peeters and Kroon (2003) propose a model that minimizes a weighted function of the number of seat shortages and cost factors, approximated by the number of carriage kilometers. They also explicitly take into account the number of changes in the compositions of the trains, since these may give an indication of the robustness of the rolling stock circulation. Their approach to deal with the restricted transition possibilities from one train composition to another is based on the concept of a *transition graph*. This concept is further explained later.

Fioole et al. (2006) deal with a more complex version of the RSCP studied by Peeters and Kroon (2003). Their problem also includes the underway combining and splitting of trains. Since the concept of transition graph is hard to apply in such cases, especially in the case that two branches of a split line do not have the same length, they use a MIP model that can be seen as an extended version of the model described by Schrijver (1993). Due to several methods to improve the quality of the model's LP relaxation, relatively large and complex instances of the problem can be solved to near optimality in an acceptable amount of time by CPLEX.

Another work on the dense network case is the one by Brucker et al. (2003), focusing on rerouting locomotive-hauled carriages. Their objective is to match supply and demand for carriages in a region in Germany, which can be considered as a dense network. The routing of the locomotives is performed at a later stage in the planning process and is not discussed in this chapter. For every trip, the departure and arrival times and stations are given as well as the regular composition of the train, i.e., the locomotive and the type and number of carriages. The rolling stock flow, imposed by the timetabled trips, is probably not feasible given the limited availability of rolling stock, i.e., the required rolling stock cannot always be available at the departure station in time, given the requirements for other trips. To obtain a match between the requested number of carriages and the available number of carriages, there exist two options:

(i) extending existing trains by coupling empty carriages to the train, and
(ii) introducing empty repositioning trips between two stations.

Obviously, the second option is much more expensive. The resulting model is a huge Integer Multicommodity Flow model that is solved heuristically using simulated annealing.

Model formulation

We present the models described by Schrijver (1993) and by Peeters and Kroon (2003). As was mentioned earlier, the first model is basically an Integer Multicommodity Flow model with several additional constraints. The underlying network is a directed time–space graph. The stations and times, corresponding to the departures and the arrivals at the stations, characterize the vertices. Two types of arcs can be distinguished, namely *trip* arcs, corresponding to a trip between two stations, and *inventory* arcs, representing the numbers of train units staying in a station between two events (i.e., a departure or arrival) at that station. These inventory arcs include an arc from the last event to the first event at every station, thereby representing the cyclicity of the circulation.

In order to further describe the model of Schrijver (1993), let M be the set of rolling stock subtypes and let $G = (V, A)$ be the time–space graph whose set of arcs is given by the union of the trip arcs, denoted by A_{T}, and the inventory arcs, denoted by A_{I}. For every trip arc $a \in A_{\mathrm{T}}$, the number of first and second class passengers is denoted by p_a^1 and p_a^2, respectively, and the maximum train length as ℓ_a. With every subtype $m \in M$, we associate the parameters q_m^1, q_m^2, c_m, and w_m, denoting, respectively, the capacities of first and second class seats, the cost and the length of a train unit of subtype $m \in M$. With every arc $a \in A$ and every subtype $m \in M$, we associate an integer variable x_a^m, representing the number of units of subtype m deployed on the trip if $a \in A_{\mathrm{T}}$, and staying in the station if $a \in A_{\mathrm{I}}$.

The objective pursued in the model is to minimize the costs of the train units deployed on the trains. To this end, the number of train units on the night inventory arcs is minimized. Letting A_{N} be the set of night inventory arcs and $\delta^+(v)$ and $\delta^-(v)$ be the sets of arcs entering and leaving vertex v, respectively, the model reads as follows:

$$\min \sum_{m \in M} \sum_{a \in A_{\mathrm{N}}} c_m x_a^m \tag{27}$$

subject to

$$\sum_{a \in \delta^-(v)} x_a^m = \sum_{a \in \delta^+(v)} x_a^m, \quad v \in V, m \in M, \tag{28}$$

$$\sum_{m \in M} q_m^k x_a^m \geqslant p_a^k, \quad a \in A_{\mathrm{T}}, k = 1, 2, \tag{29}$$

$$\sum_{m \in M} w_m x_a^m \leqslant \ell_a, \quad a \in A_{\mathrm{T}}, \tag{30}$$

$$x_a^m \geqslant 0, \text{ integer}, \quad a \in A, m \in M. \tag{31}$$

Constraints (28) are flow conservation constraints in each vertex of G. Constraints (29) impose that the capacity on each arc must not be less than the expected number of first and second class passengers. Constraints (30) impose that the sum of the lengths of the train units deployed on a trip does not exceed the maximum train length.

The model described by Peeters and Kroon (2003) can be seen as an extension of Schrijver's model. In order to deal with the fact that in each station the transition possibilities from one train composition to another one are usually limited, they use the concept of a *transition graph*. Each train has its own transition graph. As was mentioned earlier, a train is a sequence of trips to be carried out by the same rolling stock units.

In a transition graph of a train, the nodes represent the feasible train compositions on the trips, and the arcs represent the feasible transitions between compositions. More specifically, for every trip, the feasible compositions are enumerated, given the available subtypes, the limits on the train length, and the maximum allowable first and second class seat shortages. Next, all feasible transitions between the compositions on subsequent trips are determined, thereby taking care of coupling and uncoupling restrictions at the stations.

By selecting a path through the transition graph for every train, a feasible composition for every trip is determined, where also the transitions between two compositions are feasible. Since a train unit can only be coupled onto a train if the train unit is available at the right time and station, the interaction between different trains that run simultaneously is modeled by keeping track of the inventory positions of the subtypes at all relevant events at the stations during the considered time period. The inventory position at an event e of a station equals the initial allocation of the rolling stock to the station augmented with the uncoupled train units at that station and decreased with the coupled train units at that station before event e. The model ensures that, at any time and in all stations, the inventory positions of all subtypes are nonnegative. This gives rise to a Dantzig–Wolfe reformulation, whose LP relaxation is solved in Peeters and Kroon (2003) through a column generation algorithm.

Formally, let T be the set of trains, and let $\mathcal{P}^t$ be the (exponentially large) set of all paths through the transition graph of train $t \in T$. Associate a binary variable y_{tP} with every path $P \in \mathcal{P}^t$, which equals 1 if the path is selected in the solution, and 0 otherwise. The cost associated with path $P \in \mathcal{P}^t$, denoted by c_{tP}, is a weighted function of the first and second class seat shortages and the number of carriage kilometers. It equals the sum of the cost of the arcs of the path.

Letting S be the set of stations and E_s be the set of events at station s, the model also contains integer inventory variables x_{es}^m, defined for all $s \in S$, $e \in E_s$, and $m \in M$, representing the number of train units of subtype m present in station s immediately after event e. The set of events consists of the initial state 0, the departures and arrivals at the station, and a final state f. Furthermore, $q(e) \in E_s$ denotes the event at station $s \in S$ immediately preceding

event $e \in E_s$. Finally, the parameters a^m_{tPes} and b^m_{tPes} represent, respectively, the number of uncoupled and coupled units of subtype m at station s at event e for path $P \in \mathcal{P}^i$ of train $t \in T$. Recall that each path specifies the subsequent compositions of a train at the subsequent trips, so that these numbers of uncoupled and coupled train units can be derived a priori.

Let n^m be the total number of train units of subtype $m \in M$ available for the involved trains, then we can state the Dantzig–Wolfe reformulation as follows:

$$\min \sum_{t \in T} \sum_{P \in \mathcal{P}^t} c_{tP} y_{tP} \tag{32}$$

subject to

$$\sum_{P \in \mathcal{P}^t} y_{tP} = 1, \quad t \in T, \tag{33}$$

$$\sum_{s \in S} x^m_{0s} = n^m, \quad m \in M, \tag{34}$$

$$x^m_{q(e),s} + \sum_{t \in T} \sum_{P \in \mathcal{P}^t} a^m_{tPes} y_{tP} - \sum_{t \in T} \sum_{P \in \mathcal{P}^t} b^m_{tPes} y_{tP} = x^m_{es},$$
$$s \in S, m \in M, e \in E_s \setminus \{0\}, \tag{35}$$

$$x^m_{0s} = x^m_{fs}, \quad s \in S, m \in M, \tag{36}$$

$$y_{tP} \in \{0, 1\}, \quad t \in T, P \in \mathcal{P}^t, \tag{37}$$

$$x^m_{es} \geqslant 0, \text{ integer}, \quad s \in S, e \in E_s, m \in M. \tag{38}$$

Constraints (33) impose that for every train a path must be selected. Constraints (34) represent the limited availability of rolling stock, i.e., the sum of the initial inventories at all stations must be equal to the total available rolling stock. Constraints (35) are the inventory constraints, i.e., the inventory of subtype $m \in M$ after event e must equal the inventory after the previous event $q(e)$ increased with the number of uncoupled train units and decreased with the number of coupled units between $q(e)$ and e. Finally, constraints (36) imply cyclicity.

In general, this formulation has a huge number of path variables, and it would be impossible to consider them all explicitly, even for small RSCP instances. Therefore, the LP relaxation is solved using column generation, leading to a branch-and-price approach to obtain the optimal integer solution.

Experimental results

After deriving local parts of the convex hull of integral solutions of the $|M|$-dimensional polytope defined by constraints (29) and (30) for each trip arc $a \in A_T$ in a preprocessing step, real-world RSCP instances of the model of Schrijver (1993) are solved within a few seconds by CPLEX.

Table 4.
Results on real-world instances with different availabilities of rolling stock

Line	# trains	# types	# inst.	# nodes	Time
2100/15	182	2	62	11.1	6
2100/15	182	3	82	9.9	91
3000/12	115	2	45	10.9	1
3000/12	115	3	58	17.5	10

The algorithm of Peeters and Kroon (2003) for dense networks was tested on two lines of NS Reizigers, the main Dutch Train Operator. These two lines are indicated as 2100 and 3000. The experiments are carried out on a 1.6 GHz IBM NetVista 6343-25G Pentium 4 PC, using the extended LINDO/PC 6.1 optimization library for solving the LP relaxations. The problem characteristics and the results obtained are summarized in Table 4. For each of the two lines, several instances with different availabilities of rolling stock were solved. In the table, we give the number of trains in the series (*# trains*), the number of rolling stock subtypes considered (*# types*), the resulting number of instances (*# inst*), and for these instances, the average number of nodes in the branch-and-price tree (*# nodes*) and the average computing time in seconds (*time*).

5.2 Sparse network

Literature review

Since the planning horizon in a sparse network is usually much longer than in a dense network, and the same holds for the travel time to the maintenance centers, maintenance requirements must be taken explicitly into account when determining the circulation of the rolling stock.

Cordeau et al. (2001) present a model for the simultaneous locomotive and carriage assignment problem. As a result, the type and the number of both equipment types must be determined for each scheduled train, taking care that the assigned locomotives provide sufficient pulling capacity for the assigned carriages. In general, several combinations of types of locomotives and carriages are possible, where carriages typically differ in seat capacity and in class (first or second) and locomotives differ in pulling capacity. The operating speed of the different equipment types may vary, and the operating speed of a train equals the speed of its slowest component, thereby requiring some flexibility in the timetable. The solution found is cyclic and can be repeated period after period for a whole season.

Lingaya et al. (2002) present a model to deal with seasonal cycles, adapting the model of Cordeau et al. (2001) to short-term demand revisions. Based on true data of sold and requested tickets, the model tries to find alternative

cycles, seeking to maximize the expected profit, subject to several operational constraints. The locomotive cycles cannot be changed anymore, because they are the basis for the crew schedules. The most important constraints are the maintenance requirements and the minimum switching times, i.e., the minimum time needed to uncouple a carriage from a train or to couple it onto a train, between two consecutive trips that the train must make. This switching time depends on the position of the carriages in the consist. An in-body-switch is more time consuming than a switch at the tail of the train. For every carriage and day in the planning horizon on which the carriage can begin a cycle, a network is generated, representing all potential cycles for the carriage. These are based on the fixed locomotive cycles. These networks also reflect the possible initial and final conditions for the carriage, imposed by its position at the beginning of the planning horizon or by the fact that before a given day the carriage must be at the maintenance center. The model takes into account the positions of the units of equipment in the train, such that a tail switch can only be performed for the carriages positioned at the tail of the consist. A column generation procedure is proposed to solve the LP relaxation and next a heuristic branch-and-bound scheme is applied to find an integer solution. The required computing time is generally small.

Finally, Ben-Khedher et al. (1998) study the problem of allocating train units to the French TGVs. Their rolling stock allocation system is based on a capacity adjustment model that is linked to the seat reservation system and seeks to maximize the expected profit. The TGV fleet consists of several types of train units with different first and second class seat capacities. In principle, one train unit is deployed per scheduled trip. However, part of the fleet can be used to reinforce a scheduled train, because of a high number of reservations. The model is basically a huge Integer Multicommodity Flow model with side constraints that is solved with a commercial ILP solver.

Model formulation

In this section we present a simplified version of the formulation of Cordeau et al. (2001), discussing extensions in the end. Let T be the set of all trips during the time horizon, S be the set of stations, M be the set of equipment types, and $D = \{1, \ldots, n\}$ be the set of days of the time period. Then, for each scheduled trip $t \in T$, the consist is assumed to be given and the minimum number of required locomotives and carriages for the various types of equipment of the consist are denoted by n_t^m for $m \in M$. The model relies on the concept of equipment cycles, i.e., a sequence of trips and waiting times in some stations between two sequences, where each cycle starts and ends in a unique maintenance center. After a number of days, the equipment must spend some time at the maintenance center for a preventive check and possibly for some repairs. We further assume that maintenance takes place during the night and that there is enough time to do so.

We now discuss the conditions for a feasible equipment cycle. Two trips t_1 and t_2 requiring the same consist can be covered by the same equipment cycle if the arrival station of t_1 is the departure station of t_2 and there is sufficient connection time at the station. The required connection time, however, depends on whether or not carriages are coupled to or uncoupled from the train. It can occur that the same train can cover two trips only if the composition of the train is not changed at the station. To this end, the concept of *train sequence* is introduced, which is an ordered set of trips that can be covered by the same train, if the composition is not changed at an intermediate station. In addition, the connection time also depends on whether or not the train must be turned around in order to make the connection. This is the case if trips t_1 and t_2 have opposite directions.

A time–space network G^m represents all possible cycles that an equipment type $m \in M$ can make. As explained earlier, a cycle essentially consists of consecutive train sequences, on which a given equipment type can be deployed, and the waiting time that is spent at the station between two consecutive sequences. The networks are determined so as to respect the connection times. Each network has several source and sink nodes associated with, respectively, the start of the day and the end of the day at the maintenance center. For the other stations, the networks contain Start-Of-Day (SOD) and End-Of-Day (EOD) nodes for every day of the planning period. The arcs between the EOD and SOD nodes of two consecutive days allow the equipment to stay during the night at a station. The EOD nodes of the last day are connected with the SOD nodes of the first day at all stations to represent the fact that a cyclic solution that can be replicated is sought.

For each equipment type $m \in M$, let a^m be the number of available units and $\mathcal{P}^m$ be the set of possible paths, corresponding to paths from a source to a sink node in G^m. For each path $P \in \mathcal{P}^m$, c_P denotes the cost of the path (defined by the length of the path multiplied by the cost per kilometer, associated with power supply and maintenance), and parameter o_{tP} equals 1 if path P covers trip t, and 0 otherwise. Next, two parameters b_{dP} and e_{dP} are defined to represent, respectively, the start day and end day of path P. That is, b_{dP} (e_{dP}) equals 1 if path P begins (ends) on day $d \in D$. The parameter v_P equals 1 if path P crosses the end of the time horizon, i.e., the equipment stays at a station, different from the maintenance center, during the night between day p and day 1.

To formulate the problem, two sets of decision variables are defined:

(i) the flow y_P on path $P \in \mathcal{P}^m$ ($m \in M$) and
(ii) the number of units x_d^m of equipment type $m \in M$ staying at the maintenance center during day $d \in D$.

The problem can then be stated as follows:

$$\min \sum_{m \in M} \sum_{P \in \mathcal{P}^m} c_P y_P \tag{39}$$

subject to

$$\sum_{P\in\mathcal{P}^m} o_{tP} y_P \geqslant n_t^m, \quad m \in M, t \in T, \tag{40}$$

$$x_d^m + \sum_{P\in\mathcal{P}^m} e_{dP} y_P - \sum_{P\in\mathcal{P}^m} b_{d+1,P} y_P = x_{d+1}^m, \quad m \in M, d \in D \setminus \{n\}, \tag{41}$$

$$x_n^m + \sum_{P\in\mathcal{P}^m} e_{nP} y_P - \sum_{P\in\mathcal{P}^m} b_{1P} y_P = x_1^m, \quad m \in M, \tag{42}$$

$$\sum_{P\in\mathcal{P}^m} v_P y_P + \sum_{P\in\mathcal{P}^m} b_{1P} y_P + x_1^m \leqslant a^m, \quad m \in M, \tag{43}$$

$$y_P \geqslant 0, \text{ integer}, \quad m \in M, P \in \mathcal{P}^m, \tag{44}$$

$$x_d^m \geqslant 0, \text{ integer}, \quad m \in M, d \in D. \tag{45}$$

Constraints (40) take care that sufficient equipment is deployed on every trip. Constraints (41) and (42) are flow conservation constraints, i.e., they impose that the number of units of equipment type m staying at the maintenance center during day $d+1$ equals the number of units during day d increased by the number of units arriving at the maintenance center during day d and decreased by the number of units leaving on day $d+1$. Constraints (43) ensure that the total number of units of an equipment type used in the circulation does not exceed the available number a^m of equipment type m. To this end, Cordeau et al. (2001) compute the flow crossing the time horizon at each station, except at the maintenance center, and add to this flow the number of units of an equipment type that must be available at the maintenance center at the beginning of day 1, i.e., the number of units that stay at the center during day 1 and the number of units of equipment starting a cycle on day 1.

The LP relaxation of this model is solved by column generation. The column generation problem consists of finding the shortest paths through the equipment networks G^m for all $m \in M$. However, given that the network contains several source and sink nodes, and given that, depending on which source node is chosen, some sink nodes become infeasible, Cordeau et al. (2001) solve a shortest path problem for every source node, making sure that only the feasible sink nodes can be reached. Next, an integer solution is obtained by heuristically applying a truncated branch-and-bound procedure.

Cordeau et al. (2001) present several extensions of this model, namely they allow for:

(i) constraints on the locomotive pulling capacity for a sequence,
(ii) unavailability of equipment type on a given day,
(iii) storage capacity of the stations and the maintenance center,
(iv) substitution between equipment types, which implies, for example, that a second class carriage can be replaced by a first class carriage,
(v) daytime maintenance, and

(vi) various consist types for a trip.

In addition, they propose a two-phase method, where in a first phase the locomotive cycles are determined, imposing the integrality on the locomotive flows only. In a second phase the assignment of carriages is done, given the fixed locomotive cycles of the first phase and taking into account not only the circulation costs but also the switching costs, incurred if carriages are coupled to or uncoupled from the train.

The algorithm of Cordeau et al. (2001) was tested on six real-life instances of VIA Rail, a Canadian Train Operator. A weekly circulation is sought for more than 325 trips and for 130 units of equipment. There are 2 locomotive types and 4 carriage types, that can be combined into three different consist types with a different operating speed. The computing time for the first phase of the algorithm, corresponding to the determination of the locomotive cycles, lies between approximately 5000 and 50,000 seconds, depending on the instances. The second phase typically requires only a few seconds. The obtained solutions considerably improve the manual solutions of VIA Rail.

5.3 Maintenance routing

As was indicated earlier, each rolling stock unit has to visit a maintenance center regularly in order to be checked and repaired, if necessary. In this section, we describe some details of the problem of routing rolling stock units towards a maintenance center. In the same way as in the previous sections, one may distinguish here between "sparse" and "dense" systems.

In a "sparse" railway system the maintenance checks of the rolling stock are incorporated into the basic rolling stock circulation, since there is the risk that it will not be possible to get a rolling stock unit in time at the maintenance center in the moment in which this is necessary. In such a system, each rolling stock unit follows a planned cycle that starts and ends in the maintenance center. On the other hand, incorporating the maintenance checks already into the basic rolling stock plan leads to the risk that, during the operations, the maintenance plan has to be updated regularly, since, due to disruptions in the operations, the realized rolling stock circulation differs from the planned one. Another disadvantage of incorporating the maintenance checks already into the basic rolling stock plan is the fact that the length of the maintenance cycles does not fit with the cycle length of the basic rolling stock circulation.

Therefore, especially in "dense" railway systems, the rolling stock units may be routed to the maintenance center on a more or less ad hoc basis. That is, on a day-by-day basis, one determines which rolling stock units need to be taken away from the operations in order to undergo a maintenance check, and how these are routed towards the maintenance center. The latter is done preferably with a minimum number of additional train movements, since these are usually quite expensive. Rolling stock units that need to be routed towards a maintenance check are called *urgent* units. Here we focus on this routing problem of rolling stock units towards the maintenance center.

Usually, each rolling stock unit has been assigned to a series of duties. Here each duty is a set of trips that are to be carried out by the rolling stock unit. From one day to another, there may also be planned links between consecutive duties. Some series of consecutive duties pass along the maintenance center during the next days, and other series do not. Now the problem is to find appropriate *swaps* in the series of duties such that the urgent units get to serve on a series of duties passing through the maintenance center at the right time.

In order to solve the problem of efficiently routing the urgent train units to the maintenance checks, one may use an Integer Multicommodity Flow model. The underlying network is a time–space network, where the nodes correspond to the trips to be carried out. Pairs of scheduled trips in the same duty between which a swap to another duty is not possible are represented by a single node.

The planned connections from one trip to another are represented by arcs in the network. All these arcs have cost zero and capacity one. Furthermore, there are also arcs that represent the potential swaps of duties in the network. If one unit of flow passes such an arc, then this means that the corresponding rolling stock unit is swapped from one duty to another one, in order to bring it onto the right track towards the maintenance center. The costs of these arcs represent the complexity of the involved swap. It should be noted that it is usually hard to make a detailed estimate of the cost of a swap. However, the cost structure can be designed in such a way that a distinction is made between *easy* swaps, *moderately difficult* swaps, *hard* swaps, and *nearly impossible* swaps. For example, swapping the duties of two single train units that are standing at the same time at the same shunting area for more than one hour can be considered as easy. On the contrary, swapping two train units that are both the middle train unit of two trains consisting of three train units is nearly impossible, in particular if only a small amount of time overlap is available.

Each urgent unit is represented by its own commodity in the network. One unit of such a commodity is to be routed from the start of the duty that is currently served by the urgent unit to an appropriate maintenance check. Each commodity corresponding to an urgent unit has its own set of sinks. Furthermore, there is an additional commodity that represents all nonurgent units. The amount of flow of this commodity to be routed equals the number of nonurgent units. The nonurgent units have to be routed through the network in order to check that an overall feasible solution exists.

Now the problem is to find an Integer Multicommodity Flow with minimum cost in the constructed network. A solution to this problem can be interpreted as a set of node disjoint paths for the urgent train units that follow as much as possible the planned duties, and in which the complexity of the required swaps is as low as possible.

Maróti and Kroon (2007) describe a model for solving the maintenance routing problem that requires many details of the potential swaps as input. These details are required in order to be able to evaluate the complexity of the involved shunting movements. Since the required data may be hard to obtain, Maróti and Kroon (2005) also describe a simplified model, which requires

less detailed input. Since, usually, the number of urgent units that need to be routed simultaneously is small (1 up to 5), both models can be solved quickly by CPLEX.

6 Train unit shunting

Within the rush hours, the rolling stock of a passenger Train Operator is typically operating the timetable or it is in maintenance. However, outside the rush hours, and in particular during the night, most of the rolling stock is to be parked on a shunting area near one of the stations in the railway network. During the night, one of the objectives is to park the rolling stock on the shunting area in such a way that the railway operations can start up as smoothly as possible on the next morning.

The process of parking rolling stock on a shunting area, together with several related processes such as routing rolling stock between the station area and the shunting area, short term maintenance, and inside and outside cleaning, is called *shunting*. A major complicating issue is the fact that rolling stock is strongly restricted in its movements by the railway infrastructure. Therefore, units of rolling stock are easily blocking each other in their movements. In addition, time is also a restrictive resource for shunting. For example, for safety reasons, it is mandatory to respect a certain minimum headway time between any two train movements on the same track or switch.

Shunting processes are highly dependent on changes in the timetable and in the rolling stock circulation of a Train Operator: as soon as the timetable or the rolling stock circulation changes, the shunting plans have to be updated as well. Therefore, the planning of the shunting processes is the closing stone of the logistic plans of a Train Operator: tools that support planners in quickly generating shunting plans are highly relevant in practice.

In order to define the problem, we call an arriving train unit that has to be parked on a shunt yard an *arriving shunt unit*, and similarly, a train unit that has to be supplied from the shunt yard a *departing shunt unit*. Arriving shunt units are uncoupled from through trains or come from complete ending trains, and departing shunt units are units that are coupled onto through trains or form complete starting trains. Now the shunting problem can be defined as follows. Given (i) a railway station and a nearby shunting area, (ii) a timetable with, for each train, the arrival and/or departure time and platform, and its composition, and (iii) the routing costs from each platform track to each shunt track and vice versa, as well as several other cost estimates, the Train Unit Shunting Problem (TUSP) consists of matching the arriving and departing shunt units, as well as parking these shunt units on the shunt tracks, such that the total shunting costs are minimal.

The shunting costs consist of the routing costs, the train unit dependent penalties for certain shunt tracks, and the penalties for not parking shunt units that should be parked. Note that the routing costs usually vary over time,

since the (time-dependent) claims on the station infrastructure by the through trains have to be taken into account. Estimating these routing costs is a difficult problem in itself, which lies outside the scope of this chapter. For more details we refer to van't Woudt (2001). The shunting costs are mainly driven by the shunting movements. Therefore, minimizing the shunting costs can be attained by minimizing the number of shunting movements. Especially in the early morning, with the start-up of the railway operations, the number of shunting movements should be minimized. Further characteristics of the shunting problem are the following:

- Arrivals and departures of train units may be mixed in time. This implies that, within the planning horizon, the first departure may take place before the last arrival has taken place.
- Shunt units may have different types and subtypes. Train units of the same type, but possibly of different subtypes, may be combined with each other in one train. The type of a unit may restrict the set of shunt tracks where the unit can be parked.
- Shunt tracks may have different types and lengths. The type of a track determines how a unit can approach the track. Some tracks can be approached from one side only. These tracks will be called Last In First Out (LIFO) tracks. Other tracks can be approached from both sides. These tracks will be called *free* tracks.

In the matching of arriving and departing train units it is required that the matched units have the same subtype. Furthermore, if several units of different subtypes of one arriving train are matched with one departing train, then the subtypes of the units in both trains have to be in the same order. Finally, a *crossing* occurs whenever a train unit i obstructs a train unit j during the departure or arrival of train unit j. Such crossings are not allowed.

Literature review

A survey on shunting processes of cargo trains is provided by Cordeau et al. (1998). While literature on TUSP for passenger trains is scarce, there are a few references dealing with the analogous problem for mass-transit companies.

Freling et al. (2005) present a solution approach consisting of the following steps:

(i) matching of arriving and departing shunt units,
(ii) parking the shunt units on the tracks, and
(iii) routing the shunt units between the station area and the shunting area.

For several instances from the Dutch passenger Train Operator NS Reizigers, the matching step is solved quickly by CPLEX, the parking step is modeled as a Set Covering Problem and solved by dynamic column generation techniques, and the routing step is solved through a dynamic programming approach.

Tomii et al. (1999) and Tomii and Zhou (2000) propose a genetic algorithm for solving a version of the problem that takes into account several practical issues, including routing, maintenance, and duties for shunt personnel. Their problem is relatively simple, since in their context at most one train unit can be parked on a shunt track at the same time. Indeed, the latter restriction eliminates the problem of the crossings.

Di Stefano and Koci (2004) look at the problem of parking trains on the available shunting tracks in order to avoid shunting movements in the next morning. They assume that each track is long enough to host the trains assigned to it. Their main objective is to minimize the number of shunting tracks necessary to park all the trains without additional shunting movements. They consider several variants of their shunting problem, distinguished from each other by the ends of the shunting tracks that can be used for entering or leaving these tracks. For example, in the SISO-variant (Single Input Single Output), each train enters the shunting area along one end of the tracks and each train leaves the shunting area from one end of the tracks. For several variants of their problem they provide computational complexity results.

The subject of the paper by He et al. (2000) is the separation of train units from arriving trains, sorting the trains according to their destination, and finally combining them to form new departing trains. This resembles the problem of matching arriving and departing train units as described by Freling et al. (2005).

Dahlhaus et al. (2000) discuss the problem of rearranging carriages of passenger trains in a station in order to group them by destination. Their goal is to use a minimum number of tracks for this rearrangement. They show that this problem is NP-hard.

Van den Broek (2002) describes a model that can be used to test whether the capacity of the railway infrastructure around a station is sufficient for handling all the shunting movements that are enforced by the rolling stock circulation. This paper does not focus on the storage of train units, but on the scheduling of the involved shunting movements. The model assumes that the routes for the shunting movements are fixed beforehand and verifies that each shunting movement can be scheduled at a time instant such that each infra-element is occupied by at most one movement at the same time.

Blasum et al. (2000) focus on dispatching trams in a depot. They prove that this problem is NP-hard and that a restricted version of the problem can be solved in polynomial time by a dynamic programming approach. Their analysis in the context of dispatching trams is extended by Winter and Zimmermann (2000). Winter (1999) extends this approach to the case of length restrictions and mixed arrivals and departures, and presents an application at a bus depot. Furthermore, several variants of the studied problems are shown to be NP-hard.

Gallo and di Miele (2001) discuss the problem of dispatching and parking buses in a bus depot. Here the dispatching of the buses takes place in First In First Out (FIFO) order. They model this problem as a Noncrossing As-

signment Problem. They also include an extension of their model taking into account mixed arrivals and departures of buses.

Another application of dispatching and parking buses in a depot is described by Hamdouni et al. (2006). Here robust solutions are emphasized by having as little different bus types as possible in each track, and by grouping the buses of the same type as much as possible. This makes the plans less sensitive to the actual arrival times of the buses in the operations.

Model formulation

In this section we describe a model (first proposed by Schrijver (2003)) that can be used to find an appropriate allocation of train units to shunting tracks. The model focuses on storing the train units without crossings at the shunt tracks. In order to keep the presentation as clear as possible, we make the following simplifying assumptions.

- All tracks can be entered from one side only. This implies that all tracks are used in a Last In First Out (LIFO) fashion.
- Each arriving and each departing train consists of a single train unit. This simplifies the problem significantly. Indeed, if this is not assumed, then the model has to guarantee that train units from the same train are kept together as much as possible.
- No additional shunting movements related to cleaning or maintenance of train units have to be carried out. That is, briefly after the arrival of a train unit at the station, it is parked at a shunting track, and briefly before its departure from the station it leaves this shunting track again.

Relevant elements for the objective function are the routing costs between the platform zone and the shunting zone of the station, and the number of tracks on which more than one type of train units is stored. The latter is relevant for the robustness of the solution: if only one type of train units is stored on a track, then crossings will not occur at that track.

The sets A and D denote the sets of arriving and departing train units, respectively. The arriving and departing train units have been ordered according to their arrival and departure time. Each train unit has a certain type τ_a or τ_d, depending on whether the train unit is arriving or departing. An arriving train unit can be matched with a departing train unit only if they have the same type. According to these types and to the arrival and departure times, for each arriving train unit $a \in A$ there is a set $D_a \subseteq D$ of departing train units that can be matched with a. Similarly, for each departing train unit $d \in D$ we let $A_d \subseteq A$ be the set of arriving train units that can be matched with d. The length of each train unit is denoted by l_a ($a \in A$) or l_d ($d \in D$). Let T denote the set of the shunt tracks. The length of shunt track $t \in T$ is denoted by ℓ_t. The cost of routing train unit $a \in A$ or $d \in D$ to or from track $t \in T$ is denoted by c_{at} or c_{dt}, respectively.

The model contains binary variables z_{at}, equal to 1 if and only if arriving train unit $a \in A$ is parked at shunt track $t \in T$, and z_{dt}, equal to 1 if and only if departing train unit $d \in D$ is parked at shunt track $t \in T$. Moreover, binary variables x_{adt} are equal to 1 if and only if arriving train unit $a \in A$ is matched with departing train unit $d \in D$ at shunt track $t \in T$. The model reads as follows:

$$\min \sum_{t \in T} \left(\sum_{a \in A} c_{at} z_{at} + \sum_{d \in D} c_{dt} z_{dt} \right) \tag{46}$$

subject to

$$\sum_{t \in T} z_{at} = 1, \quad a \in A, \tag{47}$$

$$\sum_{t \in T} z_{dt} = 1, \quad d \in D, \tag{48}$$

$$z_{at} = \sum_{d \in D_a} x_{adt}, \quad t \in T, a \in A, \tag{49}$$

$$z_{dt} = \sum_{a \in A_d} x_{adt}, \quad t \in T, d \in D, \tag{50}$$

$$\sum_{d' \in D_a : d' > d} x_{ad't} + \sum_{a' \in A_d : a' < a} x_{a'dt} \leqslant 1,$$
$$t \in T, a \in A, d \in D, a < d, \tau_a \neq \tau_d, \tag{51}$$

$$\sum_{a' \in A : a' \leqslant a} l_{a'} z_{a't} - \sum_{d' \in D : d' < a} l_{d'} z_{d't} \leqslant \ell_t, \quad t \in T, a \in A, \tag{52}$$

$$z_{at} \in \{0, 1\}, \quad a \in A, t \in T, \tag{53}$$

$$z_{dt} \in \{0, 1\}, \quad d \in D, t \in T, \tag{54}$$

$$x_{adt} \in \{0, 1\}, \quad a \in A, d \in D, t \in T. \tag{55}$$

The objective function (46) expresses the fact that the routing costs for the shunting movements are to be minimized. (Note that also the other mentioned objective of minimizing the number of tracks with more than one train unit type can be expressed easily in the decision variables.) Constraints (47) and (48) specify that each arriving train unit and each departing train unit, respectively, is stored at a certain track. Constraints (49) specify that, if an arriving train unit a is stored at shunting track t, then it is matched there with an appropriate departing train unit d. Constraints (50) specify the same for each departing train unit d. Constraints (51) are the *crossing* constraints and guarantee that, on a single track, the arriving and departing train units are matched in a LIFO way. Indeed, constraints (51) guarantee that, if there are other train units than train unit d at track t at the moment that train unit d wants to depart from track t, then these train units have arrived earlier at track t than train unit d. Thus, train unit d is the first train unit at track t and can, thus, depart without

a crossing. Note that constraints (51) are only relevant for $\tau_a \neq \tau_d$. Indeed, $\tau_a = \tau_d$ implies $\tau'_d = \tau_a = \tau_d = \tau'_a$, and if all these train units have the same type, then they cannot create a crossing: identical train units can be exchanged if necessary. Constraints (52) guarantee that the length of each shunt track is not exceeded by the train units that are stored on the track. It is sufficient to take these constraints into account only at the arrival times of the arriving train units, as they are the only time instants at which the length of a track might be exceeded. In constraints (52), the first term represents the total length of the train units that have been stored at shunting track t before (and including) the arrival of train unit a. The second term represents the total length of the train units that have departed from track t before the arrival of train unit a. The difference between these two is the total length of the train units that are stored at track t just after the arrival of train unit a.

Model (46)–(55) can be extended to take into account also trains that consist of several train units and tracks that can be approached from both sides. Computational experiments based on station Enschede in the Netherlands show that the resulting model can be solved in an acceptable amount of time by CPLEX. For the more complex situation at station Zwolle, CPLEX usually finds high quality solutions quickly, but closing the gap between the involved lower and upper bounds may take a lot of time.

Schrijver (2003) also describes several methods for reducing the running times. One of these methods is based on the concept of so-called *virtual tracks*. This concept is based on the fact that the number of crossing constraints is quite high, and that the crossing constraints are only required for shunting tracks on which train units of different types are parked. For tracks containing train units of a single type only, the crossing constraints can be replaced by simple flow conservation constraints.

7 Crew planning

The Crew Planning Problem (CPP) is the problem to be faced by Train Operators which is concerned with building the work schedules of crews needed to cover a planned timetable. In CPP, we are given a planned timetable for the *train services* (i.e., both the actual journeys with passengers or freight, and the transfers of empty trains or equipment between different stations) to be performed every day of a certain planning horizon. Each train service has first been split into a sequence of *trips*, defined as segments of train journeys which must be serviced by the same crew (i.e., driver or conductor) without rest. Each trip is characterized by a departure time, a departure station, an arrival time, an arrival station, and possibly by additional attributes. Each daily occurrence of a trip has to be performed by one crew. In fact, each crew performs a *roster*, defined as a sequence of trips whose operational cost and feasibility depend on several rules laid down by union contracts and company regulations. The problem consists of finding a set of rosters covering every trip of the given planning horizon, so as to satisfy all the operational constraints with minimum cost.

CPP represents a very complex and challenging problem, due to both the size of the instances to be solved and the type and number of operational constraints. Typical figures for the main European Train Operator companies are a few thousand trains per day and a workforce of several thousand drivers spread among several crew depots. Usually, CPP is approached in two phases, according to the following scheme:

1. *Crew Scheduling*: The short-term schedule of the crews is considered, and a convenient set of *duties* (also called *pairings*) covering all the trips is constructed. Each duty represents a sequence of trips to be covered by a single crew within a given planning horizon overlapping at most one or two consecutive days.
2. *Crew Rostering*: The duties selected in the Crew Scheduling phase are sequenced to obtain the final rosters. Here, trips are no longer taken into account explicitly, but determine the *attributes* of the duties which are relevant for the roster feasibility and cost.

Decomposition is motivated by several reasons. First of all, each crew member is located in a given *crew depot*, which represents the starting and ending point of its work segments. A natural constraint imposes that each crew must return to its home depot within one (or two, in case an *external rest* is assigned) day, which leads to the concept of a *duty* as a short-term work segment starting and ending at the home depot and overlapping very few consecutive days. Secondly, constraints affecting the short-term work segments are different in nature from those related to the overall crew rosters. For example, the minimum time interval between two consecutive trips in a duty is a few minutes for changing trains, whereas the time interval between two consecutive duties is several hours for home rest. It is worth noting that in Crew Scheduling additional constraints, called *depot constraints*, typically impose bounds on the number of duties with given characteristics for each depot. Moreover, Crew Rostering considers each depot separately, since a roster cannot include duties of different depots.

A main objective of CPP is the minimization of the global number of crews needed to perform all the daily occurrences of the trips in the given planning horizon. In some applications, the Crew Rostering phase plays a minor role, since the corresponding constraints are rather weak and the number of crews is easily determined from the solution of the Crew Scheduling phase. This typically happens, e.g., when the considered trains cover a relatively small area, running mainly within the day, and the wide majority of the crews leave from their depot in the morning/afternoon and return back to it in the afternoon/evening. In this case, the duty performed by a crew in one day puts very limited restrictions on the duties that it may perform on the next day, and Crew Rostering is aimed at balancing the workload among the crews as evenly as possible. As a result, the objective used in the Crew Scheduling phase mainly calls for the minimization of the number of working days corresponding to the duties.

In applications involving several trains covering a wide area and/or running overnight, instead, considerable savings can be obtained through a clever sequencing of the duties obtained in the first phase. Therefore, the objective of the Crew Scheduling phase has to take into account the characteristics of the duties selected and their implication in the subsequent rostering phase. This suggests the opportunity of integrating the two phases, as we will discuss later.

Literature review

Several papers on CPP appeared in the literature. For papers concerning mass-transit and airline transportation, we refer the interested reader to the surveys in Arabeyre et al. (1969), Wren (1981), Bodin et al. (1983), Rousseau (1985), Daduna and Wren (1988), Desrochers and Rousseau (1992), Barnhart et al. (1994), Desrosiers et al. (1995), Wise (1995), Wilson (1999), Daduna and Voss (2001), and Ernst et al. (2001, 2004b), as well as to other chapters of this book.

As far as railway applications are concerned, Caprara et al. (1997) present a survey of the methods used in the literature. Most of these works focus on the Crew Scheduling phase, that is mainly solved by generating (a suitable subset of) all duties and then selecting them by solving a Set Covering Problem, possibly with additional constraints. The exact Set Covering algorithms proposed in the literature can solve instances with up to a few hundred trips and a few thousand potential duties, see Beasley (1987), Beasley and Jörnsten (1992), and Balas and Carrera (1996). At present, the best methods to solve the problem to proven optimality appear to be the state-of-the-art general-purpose ILP solvers, see Caprara et al. (2000). When larger instances are tackled, one has to resort to heuristic algorithms. Classical *greedy* algorithms are very fast in practice, but typically do not provide high quality solutions, as reported in Balas and Ho (1980) and Balas and Carrera (1996). Jacobs and Brusco (1995) and Beasley and Chu (1996) propose a genetic and a simulated annealing algorithm, respectively, whereas Lorena and Lopes (1994) use an approach based on surrogate relaxation. However, the most effective heuristic approaches to the problem appear to be those based on Lagrangian relaxation, following the seminal work by Balas and Ho (1980), and then the improvements by Beasley (1990), Fisher and Kedia (1990), Wedelin (1995), Balas and Carrera (1996), Ceria et al. (1998), Caprara et al. (1999), Kroon and Fischetti (2001), and Yagiura et al. (2006), the latter making also extensive use of local search and providing most of the best known solutions so far, even if within computing times that are considerably larger than those of, e.g., Caprara et al. (1999). Abbink et al. (2005) describe a railway crew scheduling problem in the Netherlands, where one of the objectives is to allocate the total workload as fairly as possible among the crew depots, thereby taking into account both the attractive parts of the workload and the less attractive parts.

As to Crew Rostering, the only optimization approaches for the railway case that we are aware of in the literature refer to the Italian case and are described

in Caprara et al. (1998). The other published works on the Crew Rostering Problem concern urban mass-transit systems, where the minimum number of crews required to perform the duties can easily be determined, and the objective is to evenly distribute the workload among the crews, and the airline case, for which Set Partitioning approaches can be used, see Ryan (1992), Gamache and Soumis (1998), Gamache et al. (1999), as well as the above mentioned surveys by Bodin et al. (1983), Ernst et al. (2004a, 2004b), and Wren (1981). Related cyclic staff scheduling problems are dealt with in Tien and Kamiyama (1982), Balakrishnan and Wong (1990), and Caprara et al. (2003).

The only attempts to integrate the Crew Scheduling and Rostering phases for railway CPP, to the best of our knowledge, are by Caprara et al. (2001), Ernst et al. (2001), and Freling et al. (2004). The latter two are based on Set Covering/Partitioning approaches, whereas the first one is outlined in the following sections.

Other approaches for railway CPP can be found in Morgado and Martins (1992), Chu and Chan (1998), Kwan et al. (2001), Fores et al. (2001), Freling et al. (2001), and Constantino et al. (2006).

Model formulation

Crew Scheduling and Rostering problems require finding min-cost *sequences* through a given set of *items*. Items correspond to trips for Crew Scheduling, and to duties for Crew Rostering, whereas sequences correspond to duties for Crew Scheduling, and to rosters for Crew Rostering.

A natural formulation of both problems in terms of graphs associates a node with each item, and a directed arc with each possible item transition. More specifically, one can define a directed multigraph $G = (V, A)$ having one node $j \in V$ for each item, and an arc $(i, j) \in A$ if and only if item j can appear right after item i in a feasible sequence. In some cases, two or more types of transition from item i to item j are possible, e.g., two duties in a roster may be separated by different types of rest (daily, weekly, etc.). In these cases, two or more arcs from i to j are present in G, one for each type of transition. This explains why G is a multigraph. In Crew Scheduling and Rostering problems arising in railway applications the graph G is not acyclic, since the departure and arrival times of the trips are intended modulo 24 hours. This allows an arc to connect an item i to an item j even if the end time of i is greater than the start time of j, meaning that a crew performs duties i and j on different days.

With this representation, the above problems call for a min-cost collection of *circuits* of G covering each node once, as discussed in the sequel. There are two basic ways of modeling the problem of covering the nodes of a directed graph through a suitable set of circuits as an ILP.

The first model associates a binary variable with each arc $(i, j) \in A$, indicating whether the arc is selected in an optimal solution or not. This is a natural model that is particularly suitable for cases in which the most relevant constraints concern the direct transition of the items within the sequence, hence,

they can be effectively modeled through an appropriate definition of the arc set A and the associated arc costs. On the other hand, this model can only be applied when the cost of the solution can be expressed as the sum of the costs associated with the arcs, and its LP relaxation can be very weak when the operational constraints not concerning the direct transition of two items are tight.

The second model has a possibly exponential number of binary variables, each associated with a feasible circuit of G. The main advantages of this model are that (i) it allows for circuit costs depending on the whole sequence of items, and (ii) the feasibility constraints in the model do not have to include restrictions concerning the feasibility of a single circuit. This produces a formulation whose LP relaxation is typically much tighter than in the previous model. However, the model often requires dealing with a very large number of variables. In some cases, the explicit generation of all feasible circuits is impractical, and one has to resort to a column generation approach, provided that an effective column generation procedure is available to find feasible circuits whose corresponding variable has a negative LP reduced cost.

In practice, the choice of the appropriate model and solution algorithm strongly depends on the particular structure of the problem in hand. The second model is particularly suitable for cases in which feasible circuits cover a small number of nodes, and the constraints on the circuit feasibility are cumbersome and depend on the overall node sequence. This is the situation arising in railway Crew Scheduling. On the contrary, as already mentioned, the first model appears attractive for those cases where the main feasibility constraints concern the direct sequencing of two nodes, since they can be dealt with implicitly by an appropriate definition of the arc costs. This is the case of railway Crew Rostering.

Crew Scheduling

Crew Scheduling calls for a min-cost collection of paths in G covering all the nodes (trips) once, each path satisfying a set of constraints related to the feasibility of the corresponding duty (maximum driving time, meal breaks, etc.). As already mentioned, a basic constraint for Crew Scheduling is that every duty must start and end at the crew home location (depot). It is then natural to introduce in G a dummy node d for each depot, along with the associated arcs (d, j) (respectively, (j, d)) for each node j associated with a trip which can be the first (respectively, the last) trip in a duty assigned to depot d. This allows one to convert each path representing a duty into a circuit by connecting the terminal nodes of the path to the depot node representing the home location of the crew. Let D denote the subset of nodes corresponding to the depots.

The model that is most frequently used in this case is the following. Let $\mathcal{C} = \{C_1, \ldots, C_n\}$ denote the collection of all the simple circuits of G corresponding to a feasible duty for one crew member, with $n = |\mathcal{C}|$. Each circuit C_j has an associated cost c_j, and covers the node set I_j. The binary variable y_j takes value 1 if C_j is part of the optimal solution, and 0 otherwise. Letting

$N := \{1, \ldots, n\}$, we then have the following Set Partitioning Problem with side constraints:

$$\min \sum_{j \in N} c_j y_j \tag{56}$$

subject to

$$\sum_{j \in N : v \in I_j} y_j = 1, \quad v \in V \setminus D, \tag{57}$$

$$\sum_{j \in S} y_j \leqslant |S| - 1, \quad S \in \mathcal{S}, \tag{58}$$

$$y_j \in \{0, 1\}, \quad j \in N, \tag{59}$$

where $\mathcal{S}$ denotes the family of all inclusion-minimal sets $S \subseteq N$ with the property that no feasible solution contains all circuits C_j for $j \in S$. Constraints (57) impose that each node not associated with a depot is covered by exactly one circuit, whereas inequalities (58) model the crew depot constraints.

A main advantage of the Set Partitioning model is that it allows for circuit costs depending on the whole sequence of arcs. In most cases, the feasibility constraints (58), that need not take into account restrictions concerning the feasibility of a single circuit, can be replaced by a compact set of inequalities of the form $By \leqslant w$, modeling crew depot constraints only.

Due to the nature of the services to be carried out, in most railway applications a typical crew duty covers only a small number of trips. Moreover, heavy operational constraints affect duty feasibility. This makes it sometimes practical to explicitly generate all feasible duties, which are computed and stored in a preprocessing phase called *duty generation*. In addition, operational rules allow a crew to be transported with no extra cost as a passenger on a trip, hence, the overall solution can cover a trip more than once, a main difference with respect to airline applications. In this situation, the Set Partitioning formulation (56)–(59) can profitably be replaced by its Set Covering relaxation obtained by replacing "=" with "$\geqslant$" in (57). As a result, only inclusion-maximal feasible duties, among those with the same cost, need to be considered in the duty generation. This considerably reduces the number of variables.

Most of the existing literature is focused on the pure Set Covering Problem without depot constraints, one of the exceptions being Abbink et al. (2005). The case study of this section does not take into account depot constraints. The main complication in this case is the size of the instances, with up to a few million variables and a few thousand constraints. For this case, the key ideas for the most successful heuristics by Caprara et al. (1999), Ceria et al. (1998), Wedelin (1995), and Yagiura et al. (2006) are:

(i) the use of Lagrangian relaxation combined with iterative procedures to find near-optimal multipliers (e.g., subgradient optimization), and

(ii) the use of Lagrangian costs to drive the construction of feasible solutions at each iteration of these iterative procedures.

In our case study, we used the algorithm proposed in Caprara et al. (1999), which is based on the selection of a suitable dynamic "core" problem, defined by a small subset of the variables, so as to avoid working with the full problem.

Crew Rostering

In Crew Rostering, no dummy depot nodes are needed, as all duties refer to the same crew depot. With an appropriate definition of the arc costs, the problem calls for a min-cost collection of circuits covering all the nodes once, each circuit satisfying a set of constraints related to the feasibility of the associated roster.

The same model used for the Crew Scheduling phase (and for the Crew Rostering case in some airline applications, see Gamache and Soumis, 1998; Gamache et al., 1999; Ryan, 1992) does not seem to be effective here, because of the difficulties in using a column generation technique, probably due to the combination of the relatively large number of duties in a roster and the very complicated operational constraints imposed on a roster.

An alternative model associates a binary variable x_a with each arc $a \in A$, where $x_a = 1$ if and only if arc a is used in the optimal solution. Let c_a be the cost of each arc $a \in A$. In case the objective is to minimize the overall length of the rosters (that often coincides with the number of crews needed to perform all the rosters, see Caprara et al., 1998), the cost of arc $a = (u, v)$ can be set equal to the time elapsing between the start of duty u and the start of duty v in case they are consecutive in the same roster. Moreover, let $\delta^+(v)$ and $\delta^-(v)$ represent the set of arcs of G leaving and entering node $v \in V$, respectively. The model reads:

$$\min \sum_{a \in A} c_a x_a \tag{60}$$

subject to

$$\sum_{a \in \delta^+(v)} x_a = \sum_{a \in \delta^-(v)} x_a, \quad v \in V, \tag{61}$$

$$\sum_{a \in \delta^+(v)} x_a = 1, \quad v \in V, \tag{62}$$

$$\sum_{(i,j) \in P} x_a \leqslant |P| - 1, \quad P \in \mathcal{P}, \tag{63}$$

$$x_a \in \{0, 1\}, \quad a \in A, \tag{64}$$

where the family $\mathcal{P}$ contains the inclusion-minimal arc sequences (paths or circuits) P which cannot be part of any feasible solution. Note that $|\mathcal{P}|$ generally grows exponentially with $|V|$.

Constraints (61) and (62) impose that each node is covered by exactly one circuit. Constraints (63) forbid the choice of all the arcs in any infeasible arc

subset P. Notice that $\mathcal{P}$ contains all the arc sequences which cannot be covered by a single crew because of operational constraints.

Provided that the objective function has been properly modeled by associating costs with arcs, as is the case in our case study, the main disadvantage of the model (60)–(64) is the weakness of constraints (63) when its LP relaxation is considered. (On the other hand, their very large number is in principle possible to handle, as it is easy to show that these constraints can be separated efficiently whenever a polynomial time procedure is available to tell whether a given sequence P is in $\mathcal{P}$ or not.) For this reason, it is often the case that the above model is not used directly to derive feasible solutions. Rather, heuristic algorithms can be driven by its relaxation without inequalities (63), and possibly with additional inequalities that take into account the specific structure of the problem. Their success is often related to the tightness of this relaxation. The main idea of the heuristic algorithm presented in Caprara et al. (1998), to which our case study refers, is to construct the rosters one at a time, and to choose the next duty j to be sequenced after the last duty i in the current roster as the one for which the relaxation value increment due to fixing $x_{ij} = 1$ is smallest.

An integrated approach

The main drawbacks of the classical approach that solves the Crew Scheduling and Rostering phases sequentially are the following:

- The construction of the heuristic solutions in the Crew Scheduling phase takes into account only the duty costs, and not directly the real objective function, i.e., the minimization of the global cost of the rosters for all the depots in the set D.
- The duty costs only partly reflect the constraints of the Crew Rostering phase. In particular, it is difficult to find out, a priori and separately for each depot, which are the constraints that will make the construction of the rosters for this depot difficult.
- Only one solution is kept among those found in the Crew Scheduling phase, whereas the Crew Rostering phase could produce much better rosters starting from the duties selected in some other solution which was not stored because its duty cost was not the best one.

These observations inspired the design of an integrated CPP approach, which iteratively performs the Crew Scheduling phase and, within each iteration, calls the Crew Rostering phase several times, as illustrated in Caprara et al. (2001). In the sequel, we briefly illustrate the main features of this approach.

The first novelty concerns the possible updating of the best CPP solution found so far each time a new Crew Scheduling solution is constructed. For each such solution, given by the duty set S, one computes a simple lower bound L on the value of the Crew Rostering solution for the duties in S. Letting z be the value of the best CPP solution found so far (initially, $z = \infty$), if $L < z$, one calls

the Crew Rostering phase, as the duty set S may lead to a better CPP solution. Let z^H ($\geqslant L$) be the value of the solution found by the rostering optimization phase, which is executed with a relatively small time limit. If $z^H < z$, one updates the best CPP solution obtained so far.

The second novelty concerns the definition of the duty costs for the Crew Scheduling phase. In the first application of this phase these costs are set to the same value as in the original approach (e.g., they are equal to the duty duration). This first application is run with a given time limit (with several applications of the Crew Rostering phase). At the end of the execution, the information obtained from the various calls to the Crew Rostering phase, concerning the effect of the duty characteristics on the rostering solution value (possibly depending on the associated depot), is used to update the duty costs. For instance, if overnight duties for a given depot turned out to be hard to schedule in the rostering phase, then the cost of these duties is increased. The Crew Scheduling phase is then applied again. The overall method terminates after a prefixed time limit or number of applications of the Crew Scheduling phase.

Experimental results

In this section, we illustrate the classical and integrated approaches on a set of real-world instances provided by FS Trenitalia, the main Italian Train Operator. The programs were run on a Digital Ultimate Workstation 533 MHz.

In Table 5, we report the main characteristics of each instance, namely the number of trips (*# trips*), the number of depots (*# depots*), the number of duties generated during the duty generation phase (*# duties*), and the time required by the duty generation, expressed in seconds (*time*). The first three instances refer to trains which have to be covered by crews from a single depot, namely Mestre, Milan, and Verona, respectively. The fourth instance is associated with trains to be covered by crews from the depots of Bolzano, Trieste, and Udine, whereas the last two instances refer to international trains connecting Austria, Switzerland, France, Germany, and Italy, where the various depots are located. Even if, at the moment, crews are handled separately for

Table 5.
Characteristics of the instances considered

Instance	# trips	# depots	# duties	Time
MESTRE	121	1	1,024,448	166
MILAN	502	1	874,416	629
VERONA	86	1	484,139	94
BZ_TS_UD	118	3	457,021	109
A_CH_F_D_I_1	91	13	796,771	136
A_CH_F_D_I_2	309	15	497,847	370

each country, these last instances simulate what would happen if all crews for the trains considered were handled by a unique European Train Operator.

The standard approach was run with time limits of 9000 seconds for the Crew Scheduling phase and of 1000 seconds for the Crew Rostering phase. The time limit for the first phase is much larger than the time limit for the second one since the Crew Scheduling phase, having to deal with several hundred thousand duties for these instances, is typically much more time consuming than the Crew Rostering phase, which has to deal with a few hundred duties selected in the previous phase, often subdivided among different depots. On the same instances, we also ran the integrated approach. The overall time limit was 10,000 seconds, and the time limit for each internal execution of the rostering optimization phase was set to 100 seconds. The condition for the termination of each application of the Crew Scheduling phase is actually a logical one rather than a time limit, as explained in Caprara et al. (2001).

The results are reported in Table 6. For each instance and each approach, we give the number of crews in the final CPP solution (*# crews*), along with the associated number of selected duties (*# duties*). Moreover, for the standard approach, we report a lower bound (*LB*) on the optimal value of the CPP solution computed with respect to the duties selected by the Crew Scheduling phase. For the integrated approach, we report the best overall CPP lower bound (*LB*) computed over all Set Covering solutions found during the Crew Scheduling phase. Note that this bound may not correspond to the set of duties yielding the best CPP solution. Finally, for both approaches we report the time required to obtain the best solution (*time*). This time is reported in seconds. For the standard approach, this time is equal to the sum of the times spent to find the best solutions required by the Crew Scheduling phase and, for each depot, by the Crew Rostering phase.

The table shows the considerable improvement that is achieved by the integration of the Crew Scheduling and Rostering phases, leading to an average percentage of saving of about 9.5%. Note that the time required to find the best solution is similar for the two approaches, and that the number of duties in the best solution is in some cases much smaller for the integrated approach (for

Table 6.
Solutions found with the standard and integrated approaches

Instance	Standard approach				Integrated approach			
	# crews	LB	# duties	Time	# crews	LB	# duties	Time
MESTRE	66	66	37	1118	66	66	21	148
MILAN	300	300	145	6082	288	282	152	7786
VERONA	48	48	20	102	42	42	22	344
BZ_TS_UD	66	66	30	103	60	54	28	138
A_CH_F_D_I_1	72	72	28	812	66	66	24	3434
A_CH_F_D_I_2	294	294	127	1208	246	240	94	633

instances *MESTRE* and *A_CH_F_D_I_2*). We also observe that the improvement is particularly significant for instance *A_CH_F_D_I_2*, probably because the integrated approach is able to subdivide the selected duties among the depots in a much more effective way than the standard one.

8 Perspective

In this chapter we have described several mathematical models and optimization techniques that have been developed for effectively supporting traditional planning processes in passenger railway transportation. A lot of research has been carried out in this area, both of a practical and theoretical nature. The results of this research are starting to be applied in practice. For example, several railway companies are currently using automated crew scheduling and train platforming systems. Usually, the underlying models and solution techniques of these systems are similar to the ones described in this chapter.

We have also given in the introduction a brief description of the strategic issues of long term rolling stock and crew management. Although these subjects did not yet receive a lot of attention of researchers in mathematical optimization so far, they may be fruitful areas of further research. In particular, the current solutions to these long term planning problems determine the railway systems of the future. Being able to carry out these planning processes in a rational way supported by effective decision support tools may have even more effect on the quality of future railway systems than being able to find better solutions to the operational problems in less time.

Real-time control is at the other side of the planning spectrum. The current trend in the railway industry is a shift from "planning in detail" to "effective real-time control". Disturbances and disruptions in the railway operations are inevitable. Therefore, large parts of the operational plans are never carried out. In case of a disruption, one needs as soon as possible an alternative plan. To some extent, several potential alternative plans can be prepared already, e.g., in the form of disruption scenarios for adapting the timetable and the rolling stock circulation. The latter may be particularly affective in the case of a cyclic timetable. However, crew schedules are usually noncyclic. Therefore, being able to quickly generate alternative crew schedules is highly important in case of a disruption of the railway system. In order to make this effective in practice, one needs:

(i) to have detailed information on the status quo of the railway systems (e.g., the positions of trains and crews),
(ii) to be able to quickly generate alternative crew schedules, and
(iii) to disseminate the alternative plans in a dependable way among all stake-holders.

Although, from a mathematical point of view, these problems may seem to be similar to the corresponding operational planning problems, they are quite

different, mainly due to the dynamic character of real-time control and the high time pressure.

Another fruitful area of further research is the provision of dependable dynamic travel information to the passengers, in particular in case of a disruption of the railway system. In such cases, railway companies are often blamed for providing insufficient or incorrect travel information to the passengers. Providing passengers with dependable dynamic travel information requires both quickly determining alternative travel paths for the passengers, and disseminating this information in a dependable way among the passengers.

One of the aims of this chapter is to illustrate the fact that railway systems are abundant of interesting combinatorial optimization problems. Currently, many of these problems are still solved manually in practice, partly due to the traditionally rather conservative character of the railway industry. On the other hand, the innovative possibilities provided by the effective application of mathematical models and optimization techniques were also recognized by the railway industry itself, and software applications based on these techniques recently started to be implemented. Researchers in mathematical optimization should grasp the currently available momentum and opportunities in the railway industry by not focusing too much on theoretical results, but by going for real-world applications of their models and techniques. The latter will lead to a win–win situation, both for the researchers and for the railway industry.

References

Abbink, E.W.J., van den Berg, B.W.V., Kroon, L.G., Salomon, M. (2004). Allocation of railway rolling stock for passenger trains. *Transportation Science* 38, 33–41.

Abbink, E.W.J., Fischetti, M., Kroon, L.G., Timmer, G., Vromans, M.J.C.M. (2005). Reinventing crew scheduling at Netherlands railways. *Interfaces* 35, 393–401.

Arabeyre, J., Fearnley, J., Steiger, F., Teather, W. (1969). The airline crew scheduling problem: A survey. *Transportation Science* 3, 140–163.

Balakrishnan, N., Wong, R.T. (1990). A network model for the rotating workforce scheduling problem. *Networks* 20, 25–42.

Balas, E., Carrera, M.C. (1996). A dynamic subgradient-based branch-and-bound procedure for set covering. *Operations Research* 44, 875–890.

Balas, E., Ho, A. (1980). Set covering algorithms using cutting planes, heuristics and subgradient optimization: A computational study. *Mathematical Programming Study* 12, 37–60.

Barnhart, C., Johnson, E.L., Nemhauser, G.L., Savelsbergh, M.W.P., Vance, P.H. (1994). Branch-and-price: Column generation for solving huge integer programs. In: Birge, J.R., Murty, K.G. (Eds.), *Mathematical Programming: State of the Art 1994*. University of Michigan Press, Ann Arbor, pp. 186–207.

Beasley, J.E. (1987). An algorithm for set covering problems. *European Journal of Operational Research* 31, 85–93.

Beasley, J.E. (1990). A Lagrangian heuristic for set covering problems. *Naval Research Logistics* 37, 151–164.

Beasley, J.E., Chu, P.C. (1996). A genetic algorithm for the set covering problem. *European Journal of Operational Research* 94, 392–404.

Beasley, J.E., Jörnsten, K. (1992). Enhancing an algorithm for set covering problems. *European Journal of Operational Research* 58, 293–300.

Ben-Khedher, N., Kintanar, J., Queille, C., Stripling, W. (1998). Schedule optimization at SNCF: From conception to day of departure. *Interfaces* 28, 6–23.
Billionnet, A. (2003). Using integer programming to solve the train platforming problem. *Transportation Science* 37, 213–222.
Blasum, U., Bussieck, M.R., Hochstättler, W., Moll, C., Scheel, H.H., Winter, T. (2000). Scheduling trams in the morning. *Mathematical Methods of Operations Research* 49, 137–148.
Bodin, L., Golden, B., Assad, A., Ball, M. (1983). Routing and scheduling of vehicles and crews: The state of the art. *Computers & Operations Research* 10, 63–211.
Bouma, A., Oltrogge, C. (1994). Linienplanung und Simulation für Öffentlichen Verkehrswege in Praxis und Theorie. *Eisenbahntechnische Rundschau* 43, 369–378 (in German).
Brännlund, U., Lindberg, P.O., Nöu, A., Nilsson, J.E. (1998). Railway timetabling using Lagrangian relaxation. *Transportation Science* 32, 358–369.
Brucker, J., Hurink, J.L., Rolfes, T. (2003). Routing of railway carriages: A case study. *Journal of Global Optimization* 27, 313–332.
Bussieck, M.R. (1998). Optimal lines in public rail transport. PhD thesis, TU Braunschweig.
Bussieck, M.R., Kreuzer, P., Zimmermann, U.T. (1996). Optimal lines for railway systems. *European Journal of Operational Research* 96, 54–63.
Bussieck, M.R., Winter, T., Zimmermann, U.T. (1997). Discrete optimization in public rail transport. *Mathematical Programming* 79, 415–444.
Cai, X., Goh, C.J. (1994). A fast heuristic for the train scheduling problem. *Computers & Operations Research* 21, 499–510.
Caprara, A., Fischetti, M., Toth, P., Vigo, D., Guida, P.L. (1997). Algorithms for railway crew management. *Mathematical Programming* 79, 125–141.
Caprara, A., Fischetti, M., Toth, P., Vigo, D. (1998). Modeling and solving the crew rostering problem. *Operations Research* 46, 820–830.
Caprara, A., Fischetti, M., Toth, P. (1999). A heuristic method for the set covering problem. *Operations Research* 47, 730–743.
Caprara, A., Fischetti, M., Toth, P. (2000). Algorithms for the set covering problem. *Annals of Operations Research* 98, 353–371.
Caprara, A., Monaci, M., Toth, P. (2001). A global method for crew planning in railway applications. In: Daduna, J., Voss, S. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Berlin, pp. 17–36.
Caprara, A., Fischetti, M., Toth, P. (2002). Modeling and solving the train timetabling problem. *Operations Research* 50, 851–861.
Caprara, A., Monaci, M., Toth, P. (2003). Models and algorithms for a staff scheduling problem. *Mathematical Programming* 98, 445–476.
Caprara, A., Monaci, M., Toth, P., Guida, P.L. (2006). A Lagrangian heuristic approach to real-world train timetabling problems. *Discrete Applied Mathematics* 154, 738–753.
Carey, M., Carville, S. (2003). Scheduling and platforming trains at busy complex stations. *Transportation Research* 37, 195–224.
Carey, M., Lockwood, D. (1995). A model, algorithms and strategy for train pathing. *Journal of the Operational Research Society* 46, 988–1005.
Ceria, S., Nobili, P., Sassano, A. (1998). A Lagrangian-based heuristic for large-scale set covering problems. *Mathematical Programming* 81, 215–228.
Chu, S., Chan, E. (1998). Crew scheduling of light rail transit in Hong Kong: From modeling to implementation. *Computers & Operations Research* 25, 887–894.
Claessens, M.T., van Dijk, N.M., Zwaneveld, P.J. (1998). Cost optimal allocation of passenger lines. *European Journal of Operational Research* 110, 474–489.
Constantino, A.A., de Mendonca Neto, C.F.X., Novaes, A.G. (2006). Crew rostering problem with distribution of workload based on preferences. *Annals of Operations Research*, in press.
Cordeau, J.-F., Toth, P., Vigo, D. (1998). A survey of optimization models for train routing and scheduling. *Transportation Science* 32, 380–404.
Cordeau, J.-F., Soumis, F., Desrosiers, J. (2001). Simultaneous assignment of locomotives and cars to passenger trains. *Operations Research* 49, 531–548.

Daduna, J.R., Voss, S. (Eds.) (2001). *Computer-Aided Scheduling of Public Transport*. *Lecture Notes in Economic and Mathematical Systems*, vol. 505. Springer-Verlag, Berlin.

Daduna, J.R., Wren, A. (Eds.) (1988). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economic and Mathematical Systems*, vol. 308. Springer-Verlag, Berlin.

Dahlhaus, E., Horak, P., Miller, M., Ryan, J.F. (2000). The train marshalling problem. *Discrete Applied Mathematics* 103, 41–54.

De Luca Cardillo, D., Mione, N. (1998). k l-list τ colouring of graphs. *European Journal of Operational Research* 106, 160–164.

Desrosiers, J., Dumas, Y., Solomon, M.M., Soumis, F. (1995). Time constrained routing and scheduling. In: Ball, M.O., Magnanti, T.L., Monma, C.L., Nemhauser, G.L. (Eds.), *Handbooks in Operations Research and Management Science*, vol. 8. Elsevier, Amsterdam, pp. 35–139.

Desrochers, M., Rousseau, J.M. (Eds.) (1992). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economic and Mathematical Systems*, vol. 386. Springer-Verlag, Berlin.

Dienst, H. (1978). Linienplanung in Spurgeführten Personenverkehr mit Hilfe eines Heuristischen Verfahrens, PhD thesis, TU Braunschweig (in German).

Di Stefano, G., Koci, M.L. (2004). A graph theoretical approach to the shunting problem. In: Gerards, B. (Ed.), *Proceedings of ATMOS Workshop 2003*. *Electronic Notes in Theoretical Computer Science*, vol. 92. Elsevier, Amsterdam, pp. 16–33.

Ernst, A.T., Jiang, H., Krishnamoorthy, M., Nott, H., Sier, D. (2001). An integrated optimization model for train crew management. *Annals of Operations Research* 108, 211–224.

Ernst, A.T., Jiang, H., Krishnamoorthy, M., Sier, D. (2004a). Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research* 153, 3–27.

Ernst, A.T., Jiang, H., Krishnamoorthy, M., Owens, B., Sier, D. (2004b). Annotated bibliography of personnel scheduling and rostering. *Annals of Operations Research* 127, 21–144.

Fioole, P.-J., Kroon, L.G., Maróti, G., Schrijver, A. (2006). A rolling stock circulation model for combining and splitting of passenger trains. *European Journal of Operational Research* 174, 1281–1297.

Fisher, M.L., Kedia, P. (1990). Optimal solutions of set covering/partitioning problems using dual heuristics. *Management Science* 36, 674–688.

Fores, S., Proll, L., Wren, A. (2001). Experiences with a flexible driver scheduler. In: Daduna, J., Voss, S. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Berlin, pp. 137–152.

Freling, R., Lentink, R.M., Odijk, M. (2001). Scheduling train crews: A case study for the Dutch railways. In: Daduna, J., Voss, S. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Berlin, pp. 153–166.

Freling, R., Lentink, R.M., Wagelmans, A.P.M. (2004). A decision support system for crew planning in passenger transportation using a flexible branch-and-price algorithm. *Annals of Operations Research* 127, 203–222.

Freling, R., Lentink, R.M., Kroon, L.G., Huisman, D. (2005). Shunting of passenger train units in a railway station. *Transportation Science* 39, 261–272.

Gallo, G., di Miele, F. (2001). Dispatching buses in parking depots. *Transportation Science* 35, 322–330.

Gamache, M., Soumis, F. (1998). A method for optimally solving the rostering problem. In: Yu, G. (Ed.), *Operations Research in Airline Industry*. Kluwer Academic, Boston, pp. 124–157.

Gamache, M., Soumis, F., Marquis, G., Desrosiers, J. (1999). A column generation approach for large scale aircrew rostering problems. *Operations Research* 47, 247–263.

Goossens, J.H.M., van Hoesel, C.P.M., Kroon, L.G. (2004). A branch-and-cut approach for solving railway line-planning problems. *Transportation Science* 38, 379–393.

Goossens, J.H.M., van Hoesel, C.P.M., Kroon, L.G. (2005). On solving multi-type railway line planning problems. *European Journal of Operational Research* 168, 403–424.

Hamdouni, M., Desaulniers, G., Marcotte, O., Soumis, F., Van Putten, M. (2006). Dispatching buses in a depot using block patterns. *Transportation Science* 40, 364–377.

He, S., Song, R., Chaudry, S.S. (2000). Fuzzy dispatching model and genetic algorithms for railyard operations. *European Journal of Operational Research* 124, 307–331.

Higgings, A., Kozan, E., Ferreira, L. (1997). Heuristic techniques for single line train scheduling. *Journal of Heuristics* 3, 43–62.

Huisman, D., Kroon, L.G., Lentink, R.M., Vromans, M.J.C.M. (2005). Operations research in passenger railway transportation. *Statistica Neerlandica* 59, 467–497.

Jacobs, L.W., Brusco, M.J. (1995). A local search heuristic for large set-covering problems. *Naval Research Logistics* 52, 1129–1140.

Jovanovic, D., Harker, P.T. (1991). Tactical scheduling of rail operations: The SCAN I system. *Transportation Science* 25, 46–64.

Kroon, L.G., Fischetti, M. (2001). Crew scheduling for the Netherlands railways destination: Customer. In: Daduna, J., Voss, S. (Eds.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 505. Springer-Verlag, Berlin, pp. 181–201.

Kroon, L.G., Peeters, L.W.P. (2003). A variable trip time model for cyclic railway timetabling. *Transportation Science* 37, 198–212.

Kroon, L.G., Romeijn, H.E., Zwaneveld, P.J. (1997). Routing trains through railway stations: Complexity issues. *European Journal of Operational Research* 98, 485–498.

Kroon, L.G., Dekker, R., Vromans, M.J.C.M. (2005). Cyclic railway timetabling: A stochastic optimization approach. Technical Report ERS-2005-051-LIS, Erasmus University Rotterdam.

Kwan, A., Kwan, R., Parker, M., Wren, A. (2001). Producing train driver schedules under different operating strategies. In: Wilson, N. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, Berlin, pp. 129–154.

Lindner, T. (2000). Train schedule optimization in public rail transport. PhD thesis, TU Braunschweig.

Lingaya, N., Cordeau, J.-F., Desaulniers, G., Desrosiers, J., Soumis, F. (2002). Operational car assignment at VIA rail Canada. *Transportation Research* 36, 755–778.

Lorena, L.A.N., Lopes, F.B. (1994). A surrogate heuristic for set covering problems. *European Journal of Operational Research* 79, 138–150.

Maróti, G., Kroon, L.G. (2005). Maintenance routing for train units: The transition model. *Transportation Science* 39, 518–525.

Maróti, G., Kroon, L.G. (2007). Maintenance routing for train units: The scenario model. *Computers & Operations Research* 34, 1121–1140.

Morgado, E., Martins, J. (1992). Scheduling and managing crew in the Portuguese railways. *Expert Systems with Applications* 5, 301–321.

Nachtigall, K. (1999). Periodic network optimization and fixed interval timetables. Habilitation thesis, Deutsches Zentrum für Luft-und Raumfahrt, Braunschweig.

Nachtigall, K., Voget, S. (1996). A genetic algorithm approach to periodic railway synchronization. *Computers & Operations Research* 23, 453–463.

Odijk, M. (1996). A constraint generation algorithm for the construction of periodic railway timetables. *Transportation Research* 30, 455–464.

Oliveira, E., Smith, B.M. (2000). A job-shop scheduling model for the single-track railway scheduling problem. School of Computing Research Report 2000.21, University of Leeds.

Oltrogge, C. (1994). Linienplanung für Mehrstufige Bedienungssysteme in Öffentlichen Personenverkehr. PhD thesis, TU Braunschweig (in German).

Peeters, L.W.P. (2003). Cyclic railway timetable optimization. PhD thesis, Erasmus University Rotterdam.

Peeters, M., Kroon, L.G. (2003). Circulation of railway rolling stock: A branch-and-price approach. *Computers & Operations Research*, in press.

Rousseau, J.M. (Ed.) (1985). *Computer Scheduling of Public Transport 2*. North-Holland, Amsterdam.

Ryan, D.M. (1992). The solution of massive generalized set partitioning problems in aircrew rostering. *Journal of the Operational Research Society* 43, 459–467.

Scholl, S. (2005). Customer-oriented line planning. PhD thesis, University of Kaiserslautern.

Schrijver, A. (1993). Minimum circulation of railway stock. *CWI Quarterly* 6, 205–217.

Schrijver, A. (2003). Rangeren op Opstelsporen. Technical report, CWI (in Dutch).

Schrijver, A., Steenbeek, A. (1994). Dienstregelingontwikkeling voor Railned. Technical report, CWI (in Dutch).

Serafini, P., Ukovich, W. (1989). A mathematical model for periodic event scheduling problems. *SIAM Journal on Discrete Mathematics* 2, 550–581.

Szpigel, B. (1973). Optimal train scheduling on a single track railway. In: Ross, M. (Ed.), *Operation Research'72*. North-Holland, Amsterdam, pp. 343–351.

Tien, J.M., Kamiyama, A. (1982). On manpower scheduling algorithms. *SIAM Review* 24, 275–287.

Tomii, N., Zhou, L.J. (2000). Depot shunting scheduling using combined genetic algorithms and PERT. In: Brebbia, C.A., Allan, J., Hill, R.J., Sciutto, G., Sone, S. (Eds.), *Computers in Railways VII*. *Advances in Transport*, vol. 7. WIT Press, Southampton, pp. 437–446.

Tomii, N., Zhou, L.J., Fukumara, N. (1999). Shunting scheduling problem at railway stations. In: Imam, I.F., Kodratoff, Y., El-Dessouki, A., Ali, M. (Eds.), *Lecture Notes in Artificial Intelligence*, vol. 1611. Springer-Verlag, Berlin, pp. 790–797.

van den Broek, J.J.J. (2002). Toets op Inplanbaarheid van Rangeerbewegingen. MSc thesis, Eindhoven University of Technology (in Dutch).

van't Woudt, C. (2001). Shunting of passenger train units. MSc thesis, Erasmus University Rotterdam.

Wedelin, D. (1995). An algorithm for large scale 0-1 integer programming with application to airline crew scheduling. *Annals of Operational Research* 57, 283–301.

Wilson, N. (Ed.) (1999). *Computer-Aided Transit Scheduling*. *Lecture Notes in Economic and Mathematical Systems*, vol. 471. Springer-Verlag, Berlin.

Winter, T. (1999). Online and real-time dispatching problems. PhD thesis, TU Braunschweig.

Winter, T., Zimmermann, U.T. (2000). Real-time dispatch of trams in storage yards. *Annals of Operations Research* 96, 287–315.

Wise, T.H. (1995). Column generation and polyhedral combinatorics for airline crew scheduling. PhD thesis, Cornell University.

Wren, A. (Ed.) (1981). *Computer Scheduling of Public Transport*. North-Holland, Amsterdam.

Yagiura, M., Kishida, M., Ibaraki, T. (2006). A 3-flip neighborhood local search for the set covering problem. *European Journal of Operational Research* 172, 472–499.

Zwaneveld, P.J. (1997). Railway planning and allocation of passenger lines. PhD thesis, Rotterdam School of Management.

Zwaneveld, P.J., Kroon, L.G., Romeijn, H.E., Salomon, M., Dauzère-Pérès, S., van Hoesel, C.P.M., Ambergen, H.W. (1996). Routing trains through railway stations: Model formulation and algorithms. *Transportation Science* 30, 181–194.

Zwaneveld, P.J., Kroon, L.G., van Hoesel, C.P.M. (2001). Routing trains through a railway station based on a node packing model. *European Journal of Operational Research* 128, 14–33.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14004-9

Chapter 4

Maritime Transportation

Marielle Christiansen

Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Trondheim, Norway

Department of Applied Economics and Operations Research, SINTEF Technology and Society, Trondheim, Norway
E-mail: Marielle.Christiansen@iot.ntnu.no

Kjetil Fagerholt

Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Trondheim, Norway

Department of Marine Technology, Norwegian University of Science and Technology, Trondheim, Norway

Norwegian Marine Technology Research Institute (MARINTEK), Trondheim, Norway
E-mail: Kjetil.Fagerholt@iot.ntnu.no

Bjørn Nygreen

Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, Trondheim, Norway
E-mail: Bjorn.Nygreen@iot.ntnu.no

David Ronen

College of Business Administration, University of Missouri-St. Louis, St. Louis, MO, USA
E-mail: David.Ronen@umsl.edu

1 Introduction

Maritime transportation is the major conduit of international trade, but the share of its weight borne by sea is hard to come by. The authors have surveyed the academic members of the International Association of Maritime Economists and their estimates of that elusive statistic range from 65% to 85%. Population growth, increasing standard of living, rapid industrialization, exhaustion of local resources, road congestion, and elimination of trade barriers, all of these contribute to the continuing growth in maritime transportation. In countries with long shorelines or navigable rivers, or in countries consisting of multiple islands, water transportation may play a significant role also in domestic trades, e.g., Greece, Indonesia, Japan, Norway, Philippines, and USA. Table 1 demonstrates the growth in international seaborne trade during the last couple of decades (compiled from UNCTAD, 2003, 2004).

Since 1980 the total international seaborne trade has increased by 67% in terms of weight. Tanker cargo has increased modestly, but dry bulk cargo has

Table 1.
Development of international seaborne trade (millions of tons)

Year	Tanker cargo	Dry cargo		Total
		Main bulk commodities[1]	Other	
1980	1871	796	1037	3704
1990	1755	968	1285	4008
2000	2163	1288	2421	5872
2001	2174	1331	2386	5891
2002	2129	1352	2467	5948
2003[2]	2203	1475	2490	6168

[1]Iron ore, grain, coal, bauxite/alumina, and phosphate.

[2]Estimates.

Table 2.
World fleet by vessel type (million dwt)

Year	Oil tankers	Bulk carriers	General cargo	Container ships	Other	Total
1980	339	186	116	11	31	683
1990	246	234	103	26	49	658
2000	286	281	103	69	69	808
2001	286	294	100	77	69	826
2002	304	300	97	83	60	844
2003	317	307	95	91	47	857

increased by 85%. The "Other" dry cargo, which consists of general cargo (including containerized cargo) and minor dry bulk commodities, has more than doubled.

The world maritime fleet has grown in parallel with the seaborne trade. Table 2 provides data describing the growth of the world fleet during the same period (compiled from UNCTAD, 2003, 2004).

The cargo carrying capacity of the world fleet has reached 857 million tons at the end of 2003, an increase of 25% over 1980. It is worth pointing out the fast growth in the capacity of the container ships fleet with 727% increase during the same period. These replace general cargo ships in major liner trades. To a lesser extent we see also a significant growth in the bulk carriers fleet. The gap between the increase in total trade (67%) and in the world fleet (25%) is explained by two factors. First, the boom in construction of tankers during the 1970s that resulted in excess capacity in 1980, and second, the increasing productivity of the world fleet, as demonstrated in Table 3 (compiled from UNCTAD, 2003, 2004).

Table 3.
Productivity of the world fleet

Year	World fleet (million dwt)	Total cargo* (million tons)	Total ton-miles performed (thousands of millions of ton-miles)	Tons carried per dwt	Thousands of ton-miles performed per dwt
1980	682.8	3704	16,777	5.4	25.5
1990	658.4	4008	17,121	6.1	26.0
2000	808.4	5871	23,016	7.3	28.5
2001	825.7	5840	23,241	7.1	28.1
2002	844.2	5888	23,251	7.0	27.5
2003	857.0	6168	24,589	7.2	28.7

*Inconsistencies between these data and the Total in Table 1 are in the source. However, they do not affect the productivity statistics presented in this table.

The utilization of the world fleet has increased from 5.4 tons carried per deadweight ton in 1980 to 7.2 in 2003. At the same time the annual output per deadweight ton has increased from 25.5 thousand ton-miles to 28.7.

These statistics demonstrate the dependence of the world economy on seaborne trade. A ship involves a major capital investment (usually millions of US dollars, tens of millions for larger ships) and the daily operating cost of a ship may easily amount to thousands of dollars and tens of thousands for the larger ships. Proper planning of fleets and their operations has the potential of improving their economic performance and reducing shipping costs. This is often a key challenge faced by the industry actors in order to remain competitive.

The purpose of this chapter is to introduce the reader who is familiar with Operations Research (OR), and may be acquainted with other modes of transportation, to maritime transportation. The term *maritime transportation* refers to seaborne transportation, but we shall include in this chapter also other water-borne transportation, namely inland waterways. The chapter discusses various aspects of maritime transportation operations and presents associated decision making problems and models with an emphasis on ship routing and scheduling models. This chapter focuses on prescriptive OR models and associated methodologies, rather than on descriptive models that are usually of interest to economists and public policy makers. Therefore we do not discuss statistical analysis of trade and modal-split data, nor ship safety and casualty records and related topics. To explore these topics the interested reader may refer to journals dealing with maritime economics, such as *Maritime Policy and Management* and *Maritime Economics and Logistics* (formerly *International Journal of Maritime Economics*).

The ocean shipping industry has a monopoly on transportation of large volumes of cargo among continents. Pipeline is the only transportation mode that is cheaper than ships (per cargo ton-mile) for moving large volumes of cargo over long distances. However, pipelines are far from versatile because they can

Table 4.
Comparison of operational characteristics of freight transportation modes

Operational characteristic	Mode				
	Ship	Aircraft	Truck	Train	Pipeline
Barriers to entry	small	medium	small	large	large
Industry concentration	low	medium	low	high	high
Fleet variety (physical & economic)	large	small	small	small	NA
Power unit is an integral part of the transportation unit	yes	yes	often	no	NA
Transportation unit size	fixed	fixed	usually fixed	variable	NA
Operating around the clock	usually	seldom	seldom	usually	usually
Trip (or voyage) length	days–weeks	hours–days	hours–days	hours–days	days–weeks
Operational uncertainty	larger	larger	smaller	smaller	smaller
Right of way	shared	shared	shared	dedicated	dedicated
Pays port fees	yes	yes	no	no	no
Route tolls	possible	none	possible	possible	possible
Destination change while underway	possible	no	no	no	possible
Port period spans multiple operational time windows	yes	no	no	yes	NA
Vessel-port compatibility depends on load weight	yes	seldom	no	no	NA
Multiple products shipped together	yes	no	yes	yes	NA
Returns to origin	no	no	yes	no	NA

NA – not applicable.

move only fluids in bulk over fixed routes, and they are feasible and economical only under very specific conditions. Other modes of transportation (rail, truck, air) have their advantages, but only aircraft can traverse large bodies of water, and they have limited capacity and much higher costs than ships, thus they attract high-value low-volume cargoes. Ships are probably the least regulated mode of transportation because they usually operate in international water, and very few international treaties cover their operations.

Ship fleet planning problems are different than those of other modes of transportation because ships operate under different conditions. Table 4 provides a comparison of the operational characteristics of the different freight transportation modes. We wish also to point out that ships operate mostly in international trades, which means that they are crossing multiple national jurisdictions. Actually, in many aspects aircraft are similar to ships. In both modes each unit represents a large capital investment that translates into high daily cost, both pay port fees and both operate in international routes. However, most aircraft carry mainly passengers whereas most ships haul freight. Even aircraft that transport freight carry only packaged goods whereas ships carry mostly liquid and dry bulk cargo, and often nonmixable products in separate

compartments. Since passengers do not like to fly overnight most aircraft are not operated around the clock whereas ships are operated continually. In addition, aircraft come in a small number of sizes and models whereas among ships we find a large variety of designs that result in nonhomogeneous fleets. Both ships and aircraft have higher uncertainty in their operations due to their higher dependence on weather conditions and on technology, and because they usually straddle multiple jurisdictions. However, since ships operate around the clock their schedules usually do not have buffers of planned idleness that can absorb delays. As far as trains are concerned, they have their own dedicated right of way, they cannot pass each other except for at specific locations, and their size and composition are flexible (both number of cars and number of power units). Thus the operational environment of ships is different from other modes of freight transportation, and they have different fleet planning problems.

The maritime transportation industry is highly fragmented. The web site of Lloyd's Register boasts of listing of "... over 140,000 ship and 170,000 ship owner and manager entries". In order to take advantage of differences among national tax laws, financial incentives, and operating rules, the control structure of a single vessel may involve multiple companies registered in different countries.

Although ships are the least regulated mode of transportation, there are significant legal, political, regulatory, and economic aspects involved in maritime transportation. The control structure of a ship can be designed to hide the identity of the real owner in order to minimize liability and taxes. Liability for shipping accidents may be hard to pinpoint, and damages may be impossible to collect, because numerous legal entities from different countries are usually involved, such as: owner, operator, charterer, flag of registration, shipyards, classification society, surveyors, and contractors. That is in addition to the crew that may have multiple nationalities and multiple native languages.

Only a small share of the world fleet competes directly with other modes of transportation. However, in certain situations such competition may be important and encouraged by government agencies. In short haul operations, relieving road congestion by shifting cargo and passengers to ships is often desirable and even encouraged through incentives and subsidies. A central policy objective of the European Union for the upcoming years is to improve the quality and efficiency of the European transportation system by shifting traffic to maritime and inland waterways, revitalizing the railways and linking up the different modes of transport. For further information regarding the European transport policy see the European Commission's white paper *European Transport Policy for 2010: Time to Decide* (European Commission, 2004). This source provides information about many of the European Union's programs where maritime transportation plays a prominent role.

Transportation planning has been widely discussed in the literature but most of the attention has been devoted to aircraft and road transportation by trucks and buses. Other modes of transportation, i.e., pipeline, water, and rail, have

attracted far less attention. One may wonder what the reason is for that lower attention, especially when considering the large capital investments and operating costs associated with these modes. Pipeline and rail operate over a dedicated right of way, have major barriers to entry, and relatively few operators in the market. These are some issues that may explain the lower level of attention. It is worth mentioning that research on rail planning problems has increased considerably during the last fifteen years. However, the issues mentioned for pipeline and rail do not hold for water transportation. Several explanations follow for the low attention drawn in the literature by maritime transportation planning problems:

Low visibility. In most regions people see trucks, aircraft, and trains, but not ships. Worldwide, ships are not the major transportation mode. Most cargo is moved by truck or rail. Moreover, research is often sponsored by large organizations. Numerous large organizations operate fleets of trucks, but few such organizations operate ships.

Maritime transportation planning problems are less structured. In maritime transportation planning there is a much larger variety in problem structures and operating environments. That requires customization of decision support systems, and makes them more expensive. In recent years we see more attention attracted by more complex problems in transportation planning, and this is manifested also in maritime transportation.

In maritime operations there is much more uncertainty. Ships may be delayed due to weather conditions, mechanical problems and strikes (both on board and on shore), and usually, due to their high costs, very little slack is built into their schedules. This results in a frequent need for replanning.

The ocean shipping industry has a long tradition and is fragmented. Ships have been around for thousands of years and therefore the industry may be conservative and not open to new ideas. In addition, due to the low barriers to entry there are many small, family owned, shipping companies. Most quantitative models originated in vertically integrated organizations where ocean shipping is just one component of the business.

In spite of the conditions discussed above we observe significant growth in research in maritime transportation. The first review of OR work in ship routing and scheduling appeared in 1983 (Ronen, 1983), and it traced papers back to the 1950s. A second review followed a decade later (Ronen, 1993), and recently a review of the developments over the last decade appeared (Christiansen et al., 2004). Although these reviews focused on ship routing and scheduling problems, they discussed also other related problems on all planning levels. A feature issue on OR in water transportation was published by the *European Journal of Operational Research* (Ronen, 1991), and a special issue on maritime transportation was published by *Transportation Science* (Psaraftis, 1999). A survey of decision problems that arise in container terminals is provided by Vis and de Koster (2003). The increasing research interest in OR-based maritime transportation is evidenced by the growing number of

references in the review papers. The first review paper had almost forty references covering several decades. The second one had about the same number of references most of which were from a single decade, and the most recent one has almost double that number of references for the last decade. It is worth mentioning that a large share of the research in transportation planning does not seem to be based on real cases but rather on artificially generated data. The opposite is true for maritime transportation, where the majority of problems discussed are based on real applications.

We focus our attention on planning problems in maritime transportation, and some related problems. With the fast development of commercial aircraft during the second half of the 20th century, passenger transportation by ships has diminished to ferries and cruises. Important as they are, these are small and specialized segments of maritime transportation. Therefore we shall focus here on cargo shipping. Related topics that are discussed in other chapters of this volume are excluded from this chapter, namely maritime transportation of hazardous materials (Erkut and Verter, 2007) and operations of the landside of port terminals (Crainic and Kim, 2007). We try to confine ourselves to discussion of work that is relatively easily accessible to the reader. This chapter is intended to provide a comprehensive picture, but by no means an exhaustive one.

This chapter is organized around the traditional planning levels, strategic, tactical, and operational planning. Within these planning levels we discuss the three types of operations in maritime transportation (liner, tramp, industrial) and additional specialized topics. Although we try to differentiate among the planning levels, one should remember the interplay among them. On the one hand, the higher-level or longer-term decisions set the stage for the lower-level decisions. On the other hand, one usually needs significant amount of details regarding the shorter-term decisions in order to make good longer-term decisions. We focus here on OR problems in maritime transportation, the related models, and their solution methods. Due to the fast development of computing power and memory, information regarding the computing environment becomes obsolete very quickly, and such information will only occasionally be presented.

The rest of the chapter is organized as follows: Section 2 defines terms used in OR-applications in maritime transportation and describes characteristics of the industry. Sections 3–5 are dedicated to strategic, tactical, and operational problems in maritime transportation, respectively. In these sections we present problem descriptions, models and solution approaches for the three modes of operations in maritime transportation, namely liner, industrial, and tramp. We also address in these sections naval operations, maritime supply chains, ship design and management, ship loading, contract evaluation, booking orders, speed selection, and environmental routing. The issue of robustness in maritime transportation planning is addressed in Section 6. Important trends and perspectives for the use of optimization-based decision support systems in

maritime transportation and suggestions for future research are presented in Section 7, and some concluding remarks follow in Section 8.

2 Characteristics and terminology of maritime transportation

Maritime transportation planning problems can be classified in the traditional manner according to the planning horizon into strategic, tactical and operational problems.

Among the strategic problems we find:

- market and trade selection,
- ship design,
- network and transportation system design (including the determination of transshipment points for intermodal services),
- fleet size and mix decisions (type, size, and number of vessels), and
- port/terminal location, size, and design.

The tactical problems include:

- adjustments to fleet size and mix,
- fleet deployment (assignment of specific vessels to trade routes),
- ship routing and scheduling,
- inventory ship routing,
- berth scheduling,
- crane scheduling,
- container yard management,
- container stowage planning,
- ship management, and
- distribution of empty containers.

The operational problems involve:

- cruising speed selection,
- ship loading, and
- environmental routing.

Handling of hazardous materials poses additional challenges. However, this chapter concentrates on the water-side of maritime transportation. Land-side operations and hazardous materials are discussed in other chapters in this volume. Before diving into discussion of OR models in maritime transportation it is worthwhile to take a closer look at the operational characteristics of maritime transportation and to clarify various terms that are used in this area. Figure 1 relates the demand for maritime transportation to its supply, provides a comprehensive view of these characteristics and ties them together (adapted from Jansson and Shneerson, 1987). The following three sections describe these characteristics, starting on the supply side.

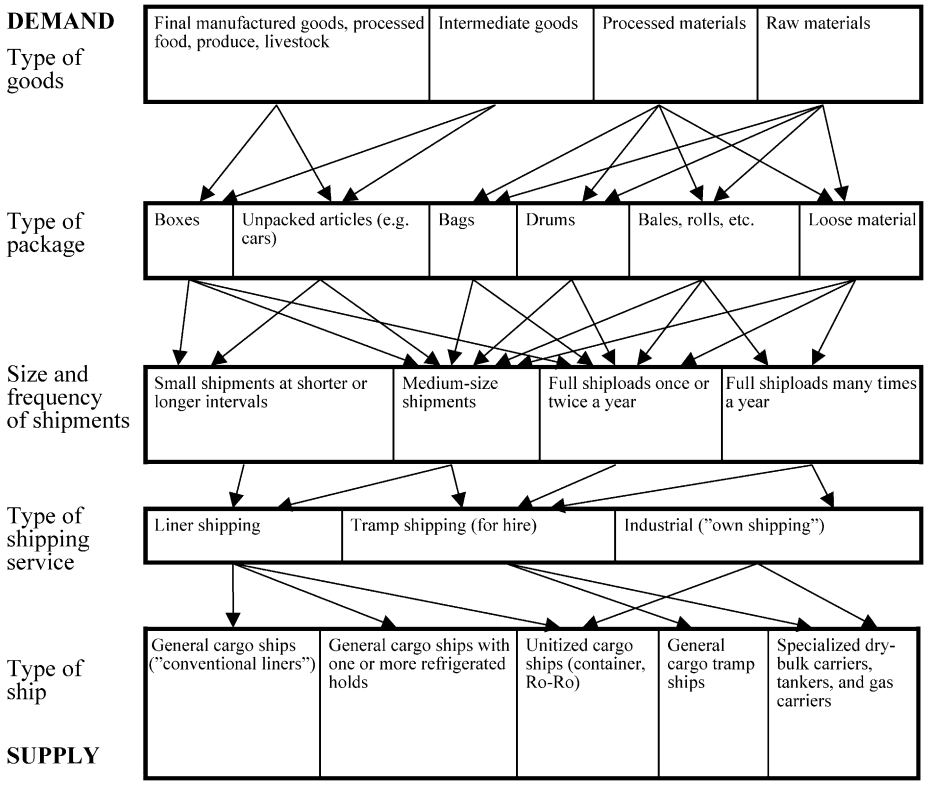


Fig. 1. Characteristic of maritime transportation demand and supply.

2.1 Ship and port characteristics

In this chapter we use the terms *ship* and *vessel* interchangeably. Although vessel may refer to other means of transportation, we shall use it in the traditional sense, referring to a ship.

Ships come in a variety of sizes. The size of a ship is measured by its weight carrying capacity and by its volume carrying capacity. Cargo with low weight per unit of volume fills the ship's volume before it reaches its weight capacity. *Deadweight* (DWT) is the weight carrying capacity of a ship, in metric tons. That includes the weight of the cargo, as well as the weight of fuels, lube oils, supplies, and anything else on the ship. *Gross Tons* (GT) is the volume of the enclosed spaces of the ship in hundreds of cubic feet.

Ships come also in a variety of types. *Tankers* are designed to carry liquids in bulk. The larger ones carry crude oil while the smaller ones usually carry oil products, chemicals, and other liquids. *Bulk carriers* carry dry bulk commodities such as iron ore, coal, grain, bauxite, alumina, phosphate, and other minerals. Some of the bulk carriers are self-discharging. They carry their own

unloading equipment, and are not dependent on port equipment for unloading their cargo. *Liquefied Gas Carriers* carry refrigerated gas under high pressure. *Container Ships* carry standardized metal containers in which packaged goods are stowed. *General Cargo* vessels carry in their holds and above deck all types of goods, usually packaged ones. These vessels often have multiple decks or floors. Since handling general cargo is labor intensive and time consuming, general cargo has been containerized during the last four decades, thus reducing the time that ships carrying such cargo spend in ports from days to hours. *Refrigerated* vessels or *reefers* are designed to carry cargo that requires refrigeration or temperature control, like fish, meat, and citrus, but can also carry general cargo. *Roll-on–Roll-off* (Ro–Ro) vessels have ramps for trucks and cars to drive on and off the vessel. Other types of vessels are *ferries, passenger ships, fishing* vessels, *service/supply* vessels, *barges* (self propelled or pushed/pulled by tugs), *research* ships, *dredgers, naval* vessels, and other, special purpose vessels. Some ships are designed as combination of the above types, e.g., ore-bulk-oil, general cargo with refrigerated compartments, passenger and Ro–Ro.

Ships operate between *ports*. Ports are used for loading and unloading cargo as well as for loading fuel, fresh water, and supplies, and discharging waste. Ports impose physical limitations on the dimensions of the ships that may call in them (ship draft, length and width), and usually charge fees for their services. Sometimes ports are used for transshipment of cargo among ships, especially when the cargo is containerized. Major container lines often operate large vessels between hub ports, and use smaller vessels to feed containers to/from spoke ports.

2.2 Types of shipping services

There are three basic modes of operation of commercial ships: *liner, tramp*, and *industrial* operations (Lawrence, 1972). *Liners* operate according to a published itinerary and schedule similar to a bus line, and the demand for their services depends among other things on their schedules. Liner operators usually control container and general cargo vessels. *Tramp* ships follow the available cargoes, similar to a taxicab. Often tramp ships engage in *contracts of affreightment*. These are contracts where specified quantities of cargo have to be carried between specified ports within a specific time frame for an agreed upon payment per unit of cargo. Tramp operators usually control tankers and dry bulk carriers. Both liner and tramp operators try to maximize their profits per time unit. *Industrial* operators usually own the cargoes shipped and control the vessels used to ship them. These vessels may be their own or on a time charter. Industrial operators strive to minimize the cost of shipping their cargoes. Such operations abound in high volume liquid and dry bulk trades of vertically integrated companies, such as: oil, chemicals, and ores. When any type of operator faces insufficient fleet capacity the operator may be able to charter in additional vessels. Whereas liners and tramp operators may give up the excess demand and related income, industrial operators must ship all their

cargoes. In cases of excess fleet capacity, vessels may be chartered out (to other operators), laid-up or even scrapped. However, when liners reduce their fleet size they must reshuffle their itineraries and/or schedules, which may result in reduced service frequency or withdrawal from certain markets. In both cases revenues may drop. An interesting historical account of the development of liner services in the US is provided by Fleming (2002, 2003).

Industrial operators, who are usually more risk-averse and tend not to charter-out their vessels, size their fleet below their long-term needs, and complement it by short-term (time or voyage/spot) charters from the tramp segment. Seasonal variations in demand, and uncertainties regarding level of future demand, freight rates, and cost of vessels (both newbuildings and second-hand) affect the fleet size decision. However, when the trade is highly specialized (e.g., liquefied gas carriers) no tramp market exists and the industrial operator must assure sufficient shipping capacity through long-term commitments. The ease of entry into the maritime industry is manifested in the tramp market that is highly entrepreneurial. This results in long periods of oversupply of shipping capacity and the associated depressed freight rates and vessel prices. However, certain market segments, such as container lines, pose large economies of scale and are hard to enter.

Naval vessels are a different breed. Naval vessels alternate between deployment at sea and relatively lengthy port periods. The major objective in naval applications is to maximize a set of measures of effectiveness. Hughes (2002) provides an interesting personal perspective of naval OR.

2.3 Cargo characteristics

Ships carry a large variety of goods. The goods may be manufactured consumer goods, unprocessed fruits and vegetables, processed food, livestock, intermediate goods, industrial equipment, processed materials, and raw materials. These goods may come in a variety of packaging, such as: boxes, bags, drums, bales, and rolls, or may be unpackaged, or even in bulk. Sometimes cargoes are unitized into larger standardized units, such as: pallets, containers, or trailers. Generally, in order to facilitate more efficient cargo handling, goods that are shipped in larger quantities are shipped in larger handling units or in bulk. During the last several decades packaged goods that required multiple manual handlings, and were traditionally shipped by liners, have been containerized into standard containers. Containerization of such goods facilitates efficient mechanized handling of the cargo, and thus saves time and money, and also reduces pilferage. Shipping containers come in two lengths, 20 feet and 40 feet. A $20'$ container carries up to approximately 28 tons of cargo with a volume of up to 1000 cubic feet. Most containers are metal boxes with an $8' \times 8'$ cross-section, but other varieties exist, such as: refrigerated containers, open top, open side, and half height. In addition there are containers of nonstandard sizes. Large containerships can carry thousands of Twenty feet Equivalent Units (TEUs), where a $40'$ container is counted as two TEUs.

In addition, goods that are shipped in larger quantities are usually shipped more often and in larger shipment sizes. Cargoes may require shoring on the ship in order to prevent them from shifting during the passage, and may require refrigeration, controlled temperature, or special handling while on board the ship. Different goods may have different weight density, thus a ship may be full either by weight or by volume, or by another measure of capacity.

2.4 Geographical characteristics

Shipping routes may be classified according to their geographical characteristics (and the corresponding type and size of vessel used): *deep-sea*, *short-sea*, *coastal*, and *inland waterways*. Due to economies of scale in shipping larger size vessels are employed in deep-sea trades between continents whereas smaller size vessels usually operate in short-sea and coastal routes, where voyage legs are relatively short. As mentioned above, smaller containerships are used on short-sea routes that feed cargo to larger vessels that operate on long deep-sea routes. A similar picture can sometimes be observed with tankers where large crude carriers used for long routes are lightered at an off shore terminal to smaller vessels (often barges). Due to draft restrictions inland waterways are used mainly by barges. Barges are used to move cargoes between the hinterland and coastal areas, often for transshipments to/from ocean-going vessels, or to move cargoes between inland ports.

2.5 Terms used in maritime transportation planning

- *Shipping* refers to moving cargoes by ships.
- The *shipper* is the owner of the transported cargo.
- A *shipment* is a specified amount of cargo that must be shipped together from a single origin to a single destination.
- *Routing* is the assignment of a sequence of ports to a vessel. *Environmental routing* or *weather routing* is the determination of the best path in a body of water that a vessel should follow.
- *Scheduling* is assigning times (or time windows) to the various events on a ship's route.
- *Deployment* refers to the assignment of the vessels in the fleet to trade routes. The differentiation between *deployment* and *scheduling* is not always clear cut. *Deployment* is usually used when vessels are designated to perform multiple consecutive trips on the same route, and therefore is associated with liners and a longer planning horizon. Liners follow a published sailing schedule and face more stable demand. *Scheduling* does not imply allocation of vessels to specific trade routes, but rather to specific shipments, and is associated with tramp and industrial operations. Due to higher uncertainty regarding future demand in these operations, their schedules usually have a shorter planning horizon.

- A *voyage* consists of a sequence of port calls, starting with the port where the ship loads its first cargo and ending where the ship unloads its last cargo and becomes empty again. A voyage may include multiple loading ports and multiple unloading ports. Liners may not become empty between consecutive voyages, and in that case a voyage starts at the port specified by the ship operator (usually a primary loading port).

Throughout this chapter we use also the following definitions:

- A *cargo* is a set of goods shipped together from a single origin to a single destination. In the vehicle routing literature it is often referred to as an order. The terms *shipment* and *cargo* are used interchangeably.
- A *load* is the set of cargoes that is on the ship at any given point in time.
- A load is considered a *full shipload* when it consists of a single cargo that for practical and/or contractual reasons cannot be carried with other cargoes.
- A *product* is a set of goods that can be stowed together in the same compartment. In the vehicle routing literature it is sometimes referred to as a commodity.
- A *loading port* is a pickup location (corresponds to a pickup node).
- An *unloading port* is a delivery location (corresponds to a delivery node).

3 Strategic planning in maritime transportation

Strategic decisions are long-term decisions that set the stage for tactical and operational ones. In maritime transportation strategic decisions cover a wide spectrum, from the design of the transportation services to accepting long-term contracts. Most of the strategic decisions are on the supply side, and these are: market selection, fleet size and mix, transportation system/service network design, maritime supply chain/maritime logistic system design, and ship design. Due to characteristics discussed earlier maritime transportation markets are usually competitive and highly volatile over time, and that complicates strategic decisions.

In this section we address the various types of strategic decisions in maritime transportation and present models for making such decisions. Section 3.1 that discusses ship design is followed by Section 3.2 that deals with fleet size and mix decisions. Section 3.3 treats network design in liner shipping, and Section 3.4 handles transportation system design. Finally, Section 3.5 addresses evaluation of long-term contracts.

In order to be able to make strategic decisions one usually needs some tactical or even operational information. Thus there is a significant overlap between strategic and tactical/operational decisions. Models used for fleet size and mix decisions and network design decisions often require evaluation of ship routing strategies. Such routing models usually fall into one of two categories, arc

flow models or path flow models. In *arc flow models* a binary variable is used to represent whether a specific vessel v travels directly from port (or customer) i to port (or customer) j. The model constructs the routes that will be used by the vessels, and the model has to keep track of both travel time and load on each vessel. In *path flow models* the routes are predefined, one way or another, and a binary variable represents whether vessel v performs route r. A route is usually a full schedule for the vessel that specifies expected arrival times and load on the vessel along the route. Such a model can focus on the set of ports or customers to serve, and only feasible routes are considered.

3.1 Ship design

A ship is basically a floating plant with housing for the crew. Therefore, ship design covers a large variety of topics that are addressed by naval architects and marine engineers, and they include structural and stability issues, materials, on-board mechanical and electrical systems, cargo handling equipment, and many others. Some of these issues have direct impact on the ship's commercial viability, and we shall focus here on two such issues, ship size and speed.

The issue of the optimal size of a ship arises when one tries to determine what is the best ship for a specific trade. In this section we deal with the optimal size of a single ship regardless of other ships that may be included in the same fleet. The latter issue, the optimal size and composition of a fleet, is discussed in Section 3.2. The optimal ship size is the one that minimizes the ship operator's cost per ton of cargo on a specific trade route with a specified cargo mix. However, one should realize that in certain situations factors beyond costs may dictate the ship size.

Ships are productive and generate income at sea. Port time is a "necessary evil" for loading and unloading cargo. Significant economies of scale exist at sea where the cost per cargo ton-mile decreases with increasing the ship size. These economies stem from the capital costs of the ship (design, construction, and financing costs), from fuel consumption, and from the operating costs (crew cost, supplies, insurance, and repairs). However, at port the picture is different. Loading and unloading rates are usually determined by the land-side cargo handling equipment and available storage space. Depending on the type of cargo and whether the cargo handling is done by the land-side equipment or by the equipment on the ship (e.g., pumps, derricks), the cargo handling rate may be constant (i.e., does not depend on the size of the ship), or, for dry cargo where multiple cranes can work in parallel, the cargo handling rate may be approximately proportional to the length of the ship. Since the size of the ship is determined by its length, width, and draft, and since the proportions among these three dimensions are practically almost constant, the size of the ship is approximately proportional to the third power of its length. Therefore, in the better case, cargo-handling rates will be proportional to the $1/3$ power of the ship size. However, when the cargo is liquid bulk (e.g., oil) the cargo-handling rate may not be related to the size of the ship.

A ship represents a large capital investment that translates into a large cost per day. Port time is expensive and presents diseconomies of scale. Thus the time of port operations caps the optimal size of ship. Generally, the longer a trade route is, the larger the share of sea-days in a voyage, and the larger the optimal ship size will be. Other factors that affect the optimal ship size are the utilization of ship capacity at sea (the "trade balance"), loading and unloading rates at the ports, and the various costs associated with the ship. On certain routes there may be additional considerations that affect the size of the ship, such as required frequency of service and availability of cargo.

A ship is a long-term investment. The useful life of a ship spans 20–30 years. Thus, the optimal ship size is a long-term decision that must be based on expectations regarding future market conditions. During the life of a ship a lot of market volatility may be encountered. Freight rates may fluctuate over a wide range, and the same is true for the cost of a ship, whether it is a second hand one or a newbuilding. When freight rates are depressed they may not even cover the variable operating costs of the ship, and the owner has very few alternatives. In the short run the owner may either reduce the daily variable operating cost of the ship by slow steaming, that results in significant reduction in fuel consumption, or the owner may lay up the ship till the market improves. Laying up a ship involves a significant set-up cost to put the ship into lay up, and, eventually, to bring it back into service. However, laying up a ship significantly reduces its daily variable operating cost. When the market is depressed, owners scrap older ships. The value of a scrapped ship is determined by the weight of its steel (the "lightweight" of the ship), but when there is high supply of ships for scrap the price paid per ton of scrap drops. Occasionally, in a very depressed market, a newly built vessel may find itself in the scrapping yard without ever carrying any cargo.

In the shorter run ship size may be limited by parameters of the specific trade, such as availability of cargoes, required frequency of service, physical limitations of port facilities such as ship draft, length, or width, and available cargo handling equipment and cargo storage capacity in the ports. In the longer run many of these limitations can be relaxed if there is an economic justification to do so. In addition there are limitations of ship design and construction technology, as well as channel restrictions in canals in the selected trade routes.

The issue of long-run optimal ship size has been discussed mainly by economists. Jansson and Shneerson (1982) presented a comprehensive model for the determination of *long-run optimal ship size*. They separated the ship capacity into two components:

- the hauling capacity (the ship size times its speed), and
- the handling capacity (cargo loaded or unloaded per time unit).

This separation facilitated the division of the total shipping costs into cost per ton of cargo carried in the voyage that does not depend on the length of the voyage, and cost per time unit. These two cost components are combined into a cost model that conveys the cost of shipping a ton of cargo a given distance. The

model requires estimation of output and cost elasticities. These elasticities, combined with the route characteristics and input prices, allow estimation of the optimal size of the ship. This model requires estimation of its parameters through regression analysis. However, high shipping market volatility over time results in low reliability of such estimates. They demonstrated the use of the model by calculating the optimal size of a coal bulk carrier for a specific trade. This work also inspects the sensitivity of the optimal ship size to four route characteristics: distance, port productivity, trade balance, and fuel costs. Most of the elasticities that are necessary for this model were estimated from several datasets in their earlier work (Jansson and Shneerson, 1978). However, that work calculated a single ship size elasticity of operating costs for each ship type. In a later study, Talley et al. (1986) analyzed short-run variable costs of tankers and concluded that the ship size elasticity of operating costs may vary according to the size of the ship of the specific type.

Modern cargo handling equipment that is customized for the specific cargo results in higher loading and unloading rates, and shorter port calls. Such equipment is justified where there is a high volume of cargo. That is usually the case in major bulk trades. Garrod and Miklius (1985) showed that under such circumstances the optimal ship size becomes very large, far beyond the capacity of existing port facilities. In addition, with such large ships the frequency of shipments drops to a point where *inventory carrying costs* incurred by the shipper start playing a significant role (the shipment size is the ship capacity). When one includes the inventory costs in the determination of the optimal ship size, that size is reduced significantly. The resulting ship sizes are still much larger than existing port facilities can accommodate, and thus the main limit on ship sizes is the draft limitation of ports. However, for a higher value cargo, or for less efficient port operations, smaller vessel sizes are optimal (see, for example, Ariel, 1991). In short-sea operations competition with other modes may play a significant role. In order to compete with other modes of transportation more frequent service may be necessary. In such cases frequency and speed of service combined with cargo availability may be a determining factor in selecting the ship size.

In *liner trades*, where there are numerous shippers, multiple ports, and a wide variety of products shipped, the inclusion of the shippers' inventory costs in the determination of the optimal ship size is more complex. Jansson and Shneerson (1985) presented the initial model for this case. In addition to the costs incurred by the ship owner/operator they included the costs of inventory that are incurred by the shipper (including the cost of safety stocks). The size (and cost) of the safety stocks is a function of the frequency of sailings on the route, and that frequency is affected by the ship size and the volume of trade. Numerous assumptions regarding the trade and the costs were necessary, and the inclusion of the shippers' costs reduced very much the optimal ship size. One could argue with the assumptions of the model, but the conclusions make sense.

Whereas Jansson and Shneerson (1985) considered a continuous review inventory control system by the shippers, Pope and Talley (1988) looked at the case of a periodic review system that is more appropriate when using a (scheduled) liner service. They found that "... optimal ship size is highly sensitive to the inventory management model selected, the treatment of stockouts and safety stocks, and the inventory management cost structure that prevails", and concluded that "rather than computing optimal ship size, it may be more appropriate to compute the optimal load size". As far as liner operations are concerned we agree with this conclusion. The optimal ship size is a long-term decision of the ship owner/operator who serves a large number of shippers. Each shipper may face different circumstances that may change over time, and therefore should be concerned with the optimal load (shipment) size. The optimal load size is a short-term decision that may change with the changing circumstances.

A historical perspective on the development of size, speed, and other characteristics of *large container ships* is provided by Gillman (1999). Cullinane and Khanna (1999) present a more recent detailed study of the economies of scale of large container ships. They take into account the considerable increase in port productivity, and take a closer look at the time in port. They find smaller diseconomies of scale (in port) than earlier studies, and show that the optimal size of a container ship continues to increase with improvements in port productivity. Taking advantage of these economies of scale to reduce shipping costs per unit while maintaining frequency of service, requires larger volumes on the trade route. This is one of the major catalysts for industry consolidation. However, McLellan (1997) injects a dose of reality to the discussion and points out that there are practical limits to the size of large containerships imposed by port draft, container handling technology, space availability, and required investments in port and transportation infrastructure.

Whereas cargo ships come in a large variety of sizes, from under 1000 DWT up to more than 500,000 DWT, their designed speed varies in a much narrower range. When one excludes outliers the ratio between the designed speed of a fast ship and a slow ship is about 2. The designed speed of a ship is a long-term decision that affects it's hauling capacity and is part of optimal ship size considerations. As a general rule the design speed of a ship increases by the square root of its length. This implies that the design speed is proportional to the 1/6 power of the size of the ship. This relationship was confirmed statistically by Jansson and Shneerson (1978), and more recently by Cullinane and Khanna (1999).

3.2 Fleet size and mix

One of the main strategic issues for shipping companies is the design of an optimal fleet. This deals with both the type of ships to include in the fleet, their sizes, and the number of ships of each size.

In order to support decisions concerning the optimal fleet of ships for an operator, we have to consider the underlying structure of the operational planning problem. This means that fleet size and mix models very often include routing decisions. For the various fleet size and mix problem types discussed in this section we can develop models that are based on the tactical models described thoroughly in Section 4.1. The objective of the strategic fleet size and mix problem is usually to minimize the fixed (setup) costs of the ships used and the variable operating costs of these ships. In a tactical routing and scheduling problem one usually minimizes only the operating costs of the ships. However, the routing decisions made in a strategic model can be later changed during tactical planning.

In addition, the fleet size and mix decisions have to be based on an estimate of demand for the transportation services. The demand forecast is highly uncertain, and stochastic techniques for considering the uncertain information are relevant for solving such strategic planning problems. Issues of robust planning are discussed in Section 6. In the literature, various demand patterns are considered where either the size of the cargoes or the frequency of sailing is specified.

In tramp shipping, contract evaluation and fleet size issues are closely related. A shipping company has to find the best split between fixed long-term cargo contracts and spot cargoes. This split should be based on estimation of future prices and demand. When considering the fleet size and mix these issues should be included. This topic is further discussed in Section 3.5.

In Section 3.2.1 we describe the fleet planning problem for a homogeneous fleet where all the vessels are of the same type, size, and cost, while the fleet size and mix for a heterogeneous fleet is the topic of Section 3.2.2.

3.2.1 Homogeneous fleet size

In this section, we want to focus on a simple industrial fleet size problem for a fleet consisting of ships of the same type, size, and cost. In the end of the section some comments regarding other studies are given.

In the fleet size planning problem considered here, a homogeneous fleet of ships is engaged in transportation of full shipload cargoes from loading ports to unloading ports. This means that just one cargo is onboard a ship at a time, and each cargo is transported directly from its loading port to it's corresponding unloading port.

All the required ship arrival times at the loading ports are fixed and known. Further, we also assume that the loading times and sailing times are known, such that the arrival times at the unloading ports can be easily calculated. The unloading times and the sailing time from each unloading port to all loading ports are also known.

The demand is such that all cargoes, given by specified loading and unloading ports, have to be serviced. The ships should be routed from the unloading ports to the loading ports in a way that minimizes the total cost of their ballast legs. Since the fleet is homogeneous and all cargoes must be transported, the

cost of the loaded legs is constant and we can leave it out. In addition, we want to minimize the number of necessary ships, and we assume that the number of ships needed dominates the sailing costs.

In the mathematical description of the problem, let $\mathcal{N}$ be the set of cargoes indexed by i and j. Cargo i is represented by a node in the network, and this node includes one loading port and one unloading port for cargo i. Since we have full information about activity times, we can determine the feasible cargo pairs (i, j). If cargo i can be serviced just before cargo j by the same ship, such an (i, j)-pair is feasible and represents an arc in the network. However, if the time between the loads is too long, the arc may be eliminated since using such arcs would result in unacceptable high waiting times. Similarly, if the departing time at node i plus the sailing time to j is greater than the given arrival time at j there will be no arc connecting the two cargoes. Let $\mathcal{N}_i^-$ and $\mathcal{N}_i^+$ be the set of all cargoes a ship can service immediately before and after servicing cargo i, respectively. Further, let $\mathcal{V}$ be the set of ships in the fleet indexed by v, and this set includes an assumption on the upper bound on the number of ships necessary. For each possible ship, we define an artificial origin cargo $o(v)$ and an artificial destination cargo $d(v)$.

The operational cost of sailing from the unloading port for cargo i to the loading port of cargo j is denoted by C_{ij}.

In the mathematical formulation, we use the following types of variables: the binary flow variable x_{ij}, $i \in \mathcal{N}$, $j \in \mathcal{N}_i^+$, equals 1, if a ship services cargo i just before cargo j, and 0 otherwise. In addition, we define flow variables for the artificial origin and artificial destination cargoes: $x_{o(v)j}$, $v \in \mathcal{V}$, $j \in \mathcal{N} \cup \{d(v)\}$, and $x_{id(v)}$, $v \in \mathcal{V}$, $i \in \mathcal{N} \cup \{o(v)\}$. If a ship v is not operating, then $x_{o(v)d(v)} = 1$.

The arc flow formulation of the industrial ship fleet size problem for one type of ships and full ship loads is as follows:

$$\min\left[\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}_i^+} C_{ij}x_{ij} - \sum_{v\in\mathcal{V}} x_{o(v)d(v)}\right] \tag{3.1}$$

subject to

$$\sum_{j\in\mathcal{N}\cup\{d(v)\}} x_{o(v)j} = 1, \quad \forall v \in \mathcal{V}, \tag{3.2}$$

$$\sum_{i\in\mathcal{N}\cup\{o(v)\}} x_{id(v)} = 1, \quad \forall v \in \mathcal{V}, \tag{3.3}$$

$$\sum_{j\in\mathcal{N}_i^+} x_{ij} + \sum_{v\in\mathcal{V}} x_{id(v)} = 1, \quad \forall i \in \mathcal{N}, \tag{3.4}$$

$$\sum_{i\in\mathcal{N}_j^-} x_{ij} + \sum_{v\in\mathcal{V}} x_{o(v)j} = 1, \quad \forall j \in \mathcal{N}, \tag{3.5}$$

$$x_{ij} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N} \cup \{o(v)\}, j \in \mathcal{N}_i^+ \cup \{d(v)\}. \tag{3.6}$$

In the first term of the objective function (3.1), we minimize the costs of the ballast legs of the ships. Since $x_{o(v)d(v)} = 1$ if ship v is not operating, the second term in the objective function minimizes the number of ships in operation. The first term is scaled in a manner that its absolute value is less than one. This means that the objective (3.1) first minimizes the number of ships in use and then as a second goal minimizes the operating costs of the ships. The second term in the objective function could easily be incorporated in the first term. However, the present form of the objective function is chosen to highlight the twofold objective. Constraints (3.2) ensure that each ship leaves its artificial origin cargo and either services one of the real cargoes or sails directly to its artificial destination cargo. In constraints (3.3) each ship in the end of its route has to arrive at its artificial destination cargo from somewhere. Constraints (3.4) ensure that the ship that services cargo i has to either service another cargo afterward or sail to its artificial destination cargo, while constraints (3.5) say that the ship servicing cargo j has to come from somewhere. Finally, the formulation involves binary requirements (3.6) on the flow variables.

We can easily see that the formulation (3.1)–(3.6) has the same structure as an assignment problem. Therefore the integrality constraints (3.6) are not a complicating factor. The problem is easily solved by any version of the simplex method or by a special algorithm for the assignment problem.

When applying a simplex method, it would be possible to have just one common artificial origin, o, and one common artificial destination, d, cargo. Then $x_{o(v)j}$, $v \in \mathcal{V}$, $j \in \mathcal{N} \cup \{d(v)\}$, and $x_{id(v)}$, $v \in \mathcal{V}$, $i \in \mathcal{N} \cup \{o(v)\}$, can be transformed into x_{oj}, $j \in \mathcal{N} \cup \{d\}$, and x_{id}, $i \in \mathcal{N} \cup \{o\}$. While the x_{oj} and x_{id} variables remain binary the variable x_{od} becomes integer.

For some problems, some of the cargoes may have a common loading port and/or a common unloading port. If the given starting times are such that several cargoes are loaded or unloaded in the same port at the same time, we assume that if this has any effect on the (un)loading times it is already accounted for in the specified data.

In a case with the same starting times in the same ports, we might change the formulation slightly. Constraints (3.4) can be considered as the constraints for leaving the unloading port for cargo i, and (3.5) as the constraints for arriving at the loading port for cargo j. We can then aggregate constraints for cargoes with the same ports and starting times. This will give more variables at the left-hand side of the constraints and a right-hand side equal to the number of aggregated constraints. The corresponding flow variables from and to the artificial cargoes will become integers rather than binary.

If some of the cargoes have the same loading and unloading ports and the same starting times then we can switch from indexing the variables by cargo numbers to indexing them by loading port, unloading port, and both loading and unloading times. Then the variables can be integer rather than binary, and their number will be reduced. Dantzig and Fulkerson (1954) pioneered such a model using a different notation for a problem with naval fuel oil tankers.

They solved a problem with 20 cargoes by using the transportation model. The number of ships was minimized and 6 ships were needed.

Later Bellmore (1968) modified the problem. An insufficient number of tankers and a utility associated with each cargo were assumed. The problem was to determine the schedules for the fleet that maximized the sum of the utilities of the carried cargoes, and it was shown to be equivalent to a transshipment problem.

Another homogeneous fleet size problem is considered in Jaikumar and Solomon (1987). Their objective is to minimize the number of tugs required to transport a given number of barges between different ports in a river system. They take advantage of the fact that the service times are negligible compared with the transit times, and of the geographical structure of the port locations on the river, and develop a highly effective polynomial exact algorithm. This problem has a line (or tree) structure, and this fact is exploited in the model definition.

Recently Sambracos et al. (2004) addressed the fleet size issue for short-sea freight services. They investigate the introduction of small containers for coastal freight shipping in the Greek Aegean Sea from two different aspects. First, a strategic planning model is developed for determining the homogeneous fleet size under known supply and demand constraints where total fuel costs and port dues are minimized. Subsequently, the operational dimension of the problem is analyzed by introducing a vehicle routing problem formulation corresponding to the periodic needs for transportation using small containers. Many simplifying assumptions are made in this study. They conclude that a 5 % cost saving may be realized by redesigning the inter-island links.

3.2.2 Heterogeneous fleet size and mix

In this section we extend the planning problem discussed in Section 3.2.1 and include decisions about the mix of different ship sizes.

We study here one particular fleet size and mix problem, where a liner shipping company wants to serve several customers that have a demand for frequent service. The problem consists of determining the best mix of ships to serve known frequencies of demand between several origin–destination port pairs. Many feasible routes are predefined, and just some of them will be used in the optimal solution. The demand is given as a minimum required number of times each port pair has to be serviced. The underlying real problem is a pickup and delivery problem. However, with predefined routes in the model, the loading and unloading aspects are not visible but hidden in the routes. Since this is a pickup and delivery problem, the frequency demand applies to a pair of ports. The ships are heterogeneous so not all ships can sail all routes. The capacity of a ship determines, among other factors, which routes it can sail. A ship is allowed to split its time between several routes.

The planning problem consists of deciding: (1) which ships to operate and (2) which routes each ship should sail and the number of voyages along each route. The first part is a strategic fleet mix and size problem and the second

part is a tactical fleet deployment problem. Fleet deployment problems are discussed in Section 4.4. The second part is used here only to find the best solution to the first part. If the demand pattern changes later, the second part can be resolved for the then available fleet.

In the mathematical description of the problem, let $\mathcal{V}$ be the set of ships indexed by v and $\mathcal{R}_v$ the set of routes that can be sailed by ship v indexed by r. The set of origin–destination port pairs is called $\mathcal{N}$ indexed by i, and each such pair needs to be serviced at least D_i times during the planning horizon.

The cost consists of two parts. We define the cost of sailing one voyage with ship v on route r as $C_{\mathrm{V}vr}$. The fixed cost for ship v during the planning horizon is called $C_{\mathrm{F}v}$. Each voyage with ship v on route r takes $T_{\mathrm{V}vr}$ time units, and A_{ir} is equal to 1 if origin–destination port pair i is serviced on route r. The length of the planning horizon is T, and we assume that the ships are available for the whole horizon. Let U_v be an upper bound on the number of voyages ship v can sail during the planning horizon.

Here we use the following types of decision variables: u_{vr}, $v \in \mathcal{V}$, $r \in \mathcal{R}_v$, represents the number of voyages along route r with ship v during the planning horizon, and s_v, $v \in \mathcal{V}$, is equal to 1 if ship v is used.

The model for the strategic fleet size and mix problem with predefined routes can then be written as

$$\min\left[\sum_{v\in\mathcal{V}}\sum_{r\in\mathcal{R}_v} C_{\mathrm{V}vr}u_{vr} + \sum_{v\in\mathcal{V}} C_{\mathrm{F}v}s_v\right] \tag{3.7}$$

subject to

$$\sum_{r\in\mathcal{R}_v} u_{vr} - U_v s_v \leqslant 0, \quad \forall v \in \mathcal{V}, \tag{3.8}$$

$$\sum_{v\in\mathcal{V}}\sum_{r\in\mathcal{R}_v} A_{ir}u_{vr} \geqslant D_i, \quad \forall i \in \mathcal{N}, \tag{3.9}$$

$$\sum_{r\in\mathcal{R}_v} T_{\mathrm{V}vr}u_{vr} \leqslant T, \quad \forall v \in \mathcal{V}, \tag{3.10}$$

$$u_{vr} \geqslant 0 \text{ and integer}, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v, \tag{3.11}$$

$$s_v \in \{0, 1\}, \quad \forall v \in \mathcal{V}. \tag{3.12}$$

Here (3.7) is the cost of sailing the used routes together with the fixed cost of the ships in operation. Constraints (3.8) ensure that the fixed costs for the ships in operation are taken into account. Constraints (3.9) say that each port pair is serviced at least the required number of times, and constraints (3.10) ensure that each ship finishes all its routes within the planning horizon. Finally, the formulation involves integer and binary requirements on the variables.

Fagerholt and Lindstad (2000) presented this model with different notation and gave an example where the model was used to plan deliveries to Norwegian petroleum installations in the North Sea. Their problem had one loading port and seven unloading installations. They managed to pre-calculate all the

feasible routes and their integer program was solved by CPLEX. The model does not ensure that services for a given port pair are properly spaced during the planning horizon. This aspect was treated manually after the model solutions were generated. Fagerholt and Lindstad (2000) report that the model solution implemented gave annual savings of several million US dollars.

Another study regarding fleet size and mix for liner routes was done by Cho and Perakis (1996). The study was performed for a container shipping company. The type of model and solution method is similar to the one used by Fagerholt and Lindstad (2000). Xinlian et al. (2000) consider a similar problem. They present a long-term fleet planning model that aims at determining which ships should be added to the existing fleet, ship retirements, and the optimal fleet deployment plan. Another study regarding the design of an optimal fleet and the corresponding weekly routes for each ship for a liner shipping system along the Norwegian coast was presented by Fagerholt (1999). The solution method is similar to the one used by Fagerholt and Lindstad (2000). In Fagerholt (1999) the solution method handled only instances where the different ships that could be selected have the same speed. This is in contrast to the work in Fagerholt and Lindstad (2000), where the ships can have different speeds. Yet another contribution within fleet size and mix for liner shipping is given by Lane et al. (1987). They consider the problem of deciding a cost efficient fleet that meets a known demand for shipping services on a defined liner trade route. The solution method has some similarities to the approach used by Fagerholt and Lindstad (2000), but the method gives no proven optimal solution since only a subset of the feasible voyage options are selected and the user determines the combination of vessel and voyage. The method has been applied on the Australia/US West coast route. Finally, resource management for a container vessel fleet is studied by Pesenti (1995). This problem involves decisions on the purchase and use of ships in order to satisfy customers' demand. A hierarchical model for the problem has been developed, and heuristic techniques, which solve problems at different decision levels, are described.

A rather special problem regarding the size of the US destroyer fleet is described in Crary et al. (2002), which illustrates the use of quantitative methods in conjunction with expert opinion. These ideas are applied to the planning scenario for the "2015 conflict on the Korean Peninsula", one of two key scenarios the Department of Defense uses for planning.

3.3 Liner network design

On all three planning levels the challenges in liner shipping are quite different from those of tramp or industrial. Liner ships are employed on more or less fixed routes, calling regularly at many ports. In contrast to industrial or tramp ships a liner ship serves demand of many shippers simultaneously, and its published route and frequency of service attract demand. The major challenges for liners at the strategic level are the design of liner routes and the associated frequency of service, fleet size and mix decisions and contract evaluation for

long-term contracts. The fleet size and mix decisions for the major market segments, including liner operations, are discussed in Section 3.2, while contract evaluation will be treated in Section 3.5. Here we focus on the design of liner routes. We split this section into three parts, where traditional liner operations are discussed in Section 3.3.1, and the more complex hub and spoke networks are considered in Section 3.3.2. Finally, we comment upon shuttle services in Section 3.3.3.

3.3.1 Traditional liner operations

Liner routes and schedules are usually set up in a manner similar to bus schedules. Before entering a particular market a liner shipping company has to thoroughly estimate the demand, revenue and cost of servicing that market. Based on this information, the company has to design its routes and to publish a sailing schedule.

Most liner companies are transporting containers, so we use here the term container(s) instead of cargo units or cargoes. We focus here on a problem where a liner container company is going to operate several different routes among a set of ports ordered more or less along a straight line. Meaning that even if a route skips a port in a contiguous sequence of ports the ship passes fairly close to the skipped port. This is usually the situation faced by longer container lines. The demands, as upper bounds on the number of transported containers, are given between all pairs of ports. The fleet of ships is heterogeneous and the planning problem consists of designing a route for each ship in a manner that maximizes the total net revenue of the fleet. One route is constructed for each ship and the ship sails as many voyages along that route as it can during the planning horizon.

The mathematical model is based on an arc flow formulation. The ports are numbered from 1 to N, and there are some strict constraints on how the routes can be constructed. Each route must have two end ports i and j, where $1 \leqslant i < j \leqslant N$. A route then starts in i and travels outbound to ports with higher and higher number until the route reaches j, where it turns around and starts its inbound travel to ports with lower and lower number until the route ends in i. A ship with i and j as end ports, does not necessarily call at all the ports between i and j, and it does not need to visit the same ports on the outbound and inbound legs of the route. See Figure 2 for an illustration of such routes.

When a ship arrives at one of its end ports it unloads all containers that are on board before it starts loading all the containers that it should load in that port. This means that each container is loaded in its loading port and stays on board the ship while the ship either sails a part of the outbound or inbound route before it is unloaded in its unloading port.

In the mathematical description of the problem, let $\mathcal{V}$ be the set of ships indexed by v and $\mathcal{N}$ the set of linearly ordered ports indexed by i, j, k, i', or j'. In addition we need the subsets $\mathcal{N}_i^+ = \{i+1, \ldots, N\}$ of ports in the

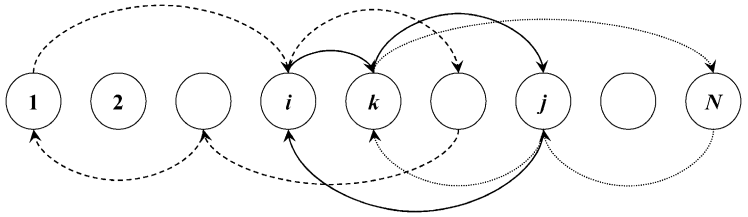


Fig. 2. Liner network design for traditional liner operations including some but not all routes.

line numbered after i and $\mathcal{N}_i^- = \{1, \ldots, i-1\}$ of ports in the line numbered before i.

The revenue for transporting one container from port i to port j is $R_{\mathrm{T}ij}$ and the cost of sailing directly from port i to port j with ship v is $C_{\mathrm{T}ijv}$. Ship v has a capacity that is measured in number of containers when it sails directly from port i to port j, and it is represented by $Q_{\mathrm{T}ijv}$. Most often it will be sufficient not to let capacity depend on the sailing leg (i, j), but in rare cases capacity may depend on weather conditions or other factors. The ship spends $T_{\mathrm{T}ijv}$ time units on that trip including the time for unloading and loading in port i. It is meaningful to assume that this time does not vary with the number of containers loaded and unloaded only if the number of such containers does not vary from call to call or that the unloading and loading time is very short compared to the sailing time. The demand as an upper bound on the number of containers transported from port i to port j during the planning horizon is denoted by $D_{\mathrm{T}ij}$. The constant S_v is the maximum time ship v is available during the planning period.

We use the following types of decision variables: e_{ijv}, $v \in \mathcal{V}$, $i \in \mathcal{N}$, $j \in \mathcal{N}$, represents the number of containers transported from port i to port j by ship v on each voyage during the planning horizon. Ship v does not necessarily sail directly from port i to port j. If ship v sails directly from port i to port j on its route, then the binary variable x_{ijv}, $v \in \mathcal{V}$, $i \in \mathcal{N}$, $j \in \mathcal{N}$, is equal to 1. The integer variable w_v, $v \in \mathcal{V}$, gives the number of whole voyages ship v manages to complete during the planning horizon. The binary variable y_{ijv}, $\forall v \in \mathcal{V}$, $i \in \mathcal{N}\backslash\{N\}$, $j \in \mathcal{N}_i^+$, is equal to 1 if ship v is allocated to a route that starts in port i and turns around in port j. These two ports i and j are called *end ports* for ship v.

A route design model for traditional liner operators can then be written as

$$\max \sum_{v\in\mathcal{V}} \sum_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} w_v (R_{\mathrm{T}ij} e_{ijv} - C_{\mathrm{T}ijv} x_{ijv}) \tag{3.13}$$

subject to

$$x_{ijv}\Bigg(\sum_{i'\in\mathcal{N}_{i+1}^-} \sum_{j'\in\mathcal{N}_{j-1}^+} e_{i'j'v} - Q_{\mathrm{T}ijv} \Bigg) \leqslant 0,$$

$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{N\}, j \in \mathcal{N}_i^+, \tag{3.14}$$

$$x_{ijv}\Bigg(\sum_{i'\in\mathcal{N}_{i-1}^{+}}\sum_{j'\in\mathcal{N}_{j+1}^{-}} e_{i'j'v} - Q_{\mathrm{T}ijv}\Bigg) \leqslant 0,$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{1\}, j \in \mathcal{N}_i^{-}, \tag{3.15}$$
$$w_v e_{ijv} \leqslant D_{\mathrm{T}ij} \sum_{j'\in\mathcal{N}_i^{+}\backslash\mathcal{N}_j^{+}} x_{ij'v},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{N\}, j \in \mathcal{N}_i^{+}, \tag{3.16}$$
$$w_v e_{ijv} \leqslant D_{\mathrm{T}ij} \sum_{j'\in\mathcal{N}_i^{-}\backslash\mathcal{N}_j^{-}} x_{ij'v},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{1\}, j \in \mathcal{N}_i^{-}, \tag{3.17}$$
$$w_v e_{ijv} \leqslant D_{\mathrm{T}ij} \sum_{i'\in\mathcal{N}_j^{-}\backslash\mathcal{N}_i^{-}} x_{i'jv},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{N\}, j \in \mathcal{N}_i^{+}, \tag{3.18}$$
$$w_v e_{ijv} \leqslant D_{\mathrm{T}ij} \sum_{i'\in\mathcal{N}_j^{+}\backslash\mathcal{N}_i^{+}} x_{i'jv},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{1\}, j \in \mathcal{N}_i^{-}, \tag{3.19}$$
$$\sum_{v\in\mathcal{V}} w_v e_{ijv} \leqslant D_{\mathrm{T}ij}, \quad \forall i \in \mathcal{N}, j \in \mathcal{N}, i \neq j, \tag{3.20}$$
$$w_v\Bigg(\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} T_{\mathrm{T}ijv} x_{ijv}\Bigg) \leqslant S_v, \quad \forall v \in \mathcal{V}, \tag{3.21}$$
$$\sum_{i\in\mathcal{N}\backslash\{N\}}\sum_{j\in\mathcal{N}_i^{+}} y_{ijv} \leqslant 1, \quad \forall v \in \mathcal{V}, \tag{3.22}$$
$$y_{ijv}\Bigg(\sum_{j'\in\mathcal{N}_i^{+}\backslash\mathcal{N}_j^{+}} x_{ij'v} - 1\Bigg) = 0,$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{N\}, j \in \mathcal{N}_i^{+}, \tag{3.23}$$
$$y_{ijv}\Bigg(\sum_{j'\in\mathcal{N}_i^{+}\backslash\mathcal{N}_j^{+}} x_{j'iv} - 1\Bigg) = 0,$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{N\}, j \in \mathcal{N}_i^{+}, \tag{3.24}$$
$$y_{ijv}\Bigg(\sum_{i'\in\mathcal{N}_k^{-}\backslash\mathcal{N}_i^{-}} x_{i'kv} - \sum_{j'\in\mathcal{N}_k^{+}\backslash\mathcal{N}_j^{+}} x_{kj'v}\Bigg) = 0,$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}\backslash\{N\}, j \in \mathcal{N}_i^{+}, k \in \mathcal{N}_i^{+}\backslash\mathcal{N}_{j-1}^{+}, \tag{3.25}$$

$$y_{ijv}\left(\sum_{i'\in\mathcal{N}_k^+\setminus\mathcal{N}_j^+} x_{i'kv} - \sum_{j'\in\mathcal{N}_k^-\setminus\mathcal{N}_i^-} x_{kj'v}\right) = 0,$$

$$\forall v \in \mathcal{V}, i \in \mathcal{N}\setminus\{N\}, j \in \mathcal{N}_i^+, k \in \mathcal{N}_i^+\setminus\mathcal{N}_{j-1}^+, \tag{3.26}$$

$$x_{ijv} \in \{0,1\}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}, j \in \mathcal{N}, i \neq j, \tag{3.27}$$

$$e_{ijv} \geqslant 0, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}, j \in \mathcal{N}, i \neq j, \tag{3.28}$$

$$w_v \geqslant 0 \text{ and integer}, \quad \forall v \in \mathcal{V}, \tag{3.29}$$

$$y_{ijv} \in \{0,1\}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}\setminus\{N\}, j \in \mathcal{N}_i^+. \tag{3.30}$$

The objective function (3.13) maximizes the difference between the revenue from transporting containers and the cost of operating the ships. The capacity of the ship might vary from leg to leg of the voyage, and (3.14) and (3.15) represent the capacity constraints for the possible outbound and inbound legs. To be able to transport containers from port i to port j on ship v, the ship needs to depart from i, either directly to j or to a port between them. In addition the ship needs to arrive in j either directly from i or from a port between them. The four constraints, (3.16)–(3.19), express these issues. The constraints for the outbound and inbound parts of the voyage had to be given separately. Each of these constraints ensures that if none of the binary flow variables, $x_{i'jv}$ or $x_{ij'v}$, is equal to 1, the number of containers transported by ship v from port i to port j during the planning horizon is zero. When the binary flow variables are equal to 1, the corresponding constraint is redundant. The demands as upper bounds on the number of transported containers are expressed in (3.20), and the upper bound on the number of voyages for each ship is expressed in (3.21). The connectivity of each route is expressed by (3.22)–(3.26). Constraints (3.22) ensure that each ship can have only one pair of end ports (one starting port i and one turning port j). A ship that starts in port i and turns around in port j, needs to leave i for a port not farther away than j and it needs to arrive in i from a port not farther away than j. This is expressed in (3.23) and (3.24). For each port, k, numbered between i and j, the same ship must arrive in k the same number of times, 0 or 1, as the number of times it departs from k, both on the outbound part and on the inbound part of the route. This is taken care of by (3.25) and (3.26). The turning around in port j is taken care of by the fact that if port k is the last port ship v visits before it reaches port j, then constraints (3.25) say that the ship has to travel directly from port k to port j. And if port k' is the first port ship v visits on the inbound part of its voyage after leaving port j, then constraints (3.26) say that the ship has to travel directly from port j to port k'.

Rana and Vickson (1988) presented a model for routing of one ship. Later (Rana and Vickson, 1991) they enhanced the model to a fleet of ships, and this latter model is the same as the one presented here with a different notation, and with constraints (3.14) and (3.15) written linearly. The solution method used by Rana and Vickson can be summarized as follows. They started with reducing the nonlinearities in the model. If we look carefully at constraints

(3.14)–(3.26) we see that constraints (3.20) are the only type of constraints that is summed over v. All the other constraints are written separately for each ship. The authors exploited this fact to apply Lagrangian relaxation to constraints (3.20). Then the problem decomposes into one problem for each ship. However, they needed to iterate or optimize over the Lagrangian multipliers. In solving the problem for each ship they solved it for different fixed values for the number of voyages. In this way, they got mixed linear integer subproblems, which they solved to near optimality by using Bender's decomposition. They give results for problems with 3 ships and between 5 and 20 ports. On average their solutions are about 2% from the upper bounds.

All the nonlinearities in (3.13)–(3.26) consist of products of two variables or one variable and a linear expression in other variables. Apart from the terms with $w_v e_{ijv}$, all the nonlinear terms consist of products where at least one variable is binary. So by first expressing w_v by binary variables, we can remove the product terms by defining one new variable and three new constraints for each product term as described by Williams (1999) in Chapter 9.2. We might then, over a decade after the publication of that paper (Rana and Vickson, 1991), be able to solve small instances of the underlying problem by using standard commercial software for mixed integer programming.

A rather special liner shipping problem is described by Hersh and Ladany (1989). However, the structure of the problem has some similarities to the problem described here. A company leasing a luxury ocean liner for Christmas cruises from Southern Florida is confronted with the problem of deciding upon the type of cruises to offer. The decision variables in the problem include the routing of the ship, the duration of the cruises, the departure dates, and the fare schedules of the cruises.

3.3.2 Hub and spoke networks

Containers are usually both faster and cheaper to load and unload than the general cargo that is stuffed in them. This means that containers can efficiently be loaded and unloaded several times between their origin and their final destination. One type of maritime transportation systems for containers is the so-called *hub and spoke network* or a *trunk line and feeder system*. In such systems we have a trunk line operating between the major ports (hubs) and a system of feeder ships working in the geographical region around each hub port visited by the trunk line. The ports feeding containers to a hub are the spokes. Thus, a container is typically loaded and unloaded three times. First a feeder ship transports the container from its initial loading port to a trunk line hub port. Then a trunk line ship transports the container to another trunk line hub port, and finally another feeder ship takes the container to its final unloading port. Such networks are further described in the chapter by Crainic and Kim (2007) on intermodal transportation in this handbook.

Here we study a short-sea application of a feeder system around one trunk line hub port with a homogeneous fleet of feeder ships. We model the transportation of containers between one hub port and a set of feeder ports (spokes)

in one geographical region. Each container is either loaded or unloaded in the hub.

The demands both to and from a spoke port are assumed to increase with the number of visits in the port during the planning horizon. These demands are upper bounds on the number of containers available for transportation, but the shipping company is not obliged to satisfy the total demand.

The planning problem consists of choosing which of a possible huge set of predefined routes to use and how many voyages to sail along the chosen routes, while maximizing the net revenue. Figure 3 illustrates the problem with one hub and several spokes. The designed routes might be overlapping.

In the mathematical description of the problem, let $\mathcal{R}$ be the set of predefined routes indexed by r and $\mathcal{N}$ be the set of ports, excluding the hub, indexed by i. Further, let $\mathcal{N}_r$ be the set of ports, excluding the hub, visited on route r. The routes that visit port i are given by the set $\mathcal{R}_i$. The ports called after port i on route r belong to the set $\mathcal{N}_{ir}^+$ and the ports called before and including port i on route r belong to the set $\mathcal{N}_{ir}^-$. Let $\mathcal{M}$ be the set of possible calls at the same port during the planning horizon indexed by m.

We assume that there are fixed revenues, $R_{\mathrm{L}i}$ and $R_{\mathrm{U}i}$, for carrying one container to and from port i. The cost consists of three parts. We call the fixed cost of operating a ship during the planning horizon C_{F}. The cost of sailing one voyage along route r is $C_{\mathrm{V}r}$ and the cost of unloading (loading) one container in port i on route r is $C_{\mathrm{U}ir}$ ($C_{\mathrm{L}ir}$). Since the fleet is homogeneous and the unit costs are specified before we know the loading pattern along the routes, we will normally have $C_{\mathrm{U}ir}$ and $C_{\mathrm{L}ir}$ independent of r. The time each ship is available during the planning horizon is called the shipping season S. The sailing time for one voyage along route r is $T_{\mathrm{V}r}$ and the capacity measured in number of containers of a ship is Q. The demand is specified in the following way: $D_{\mathrm{U}im}$

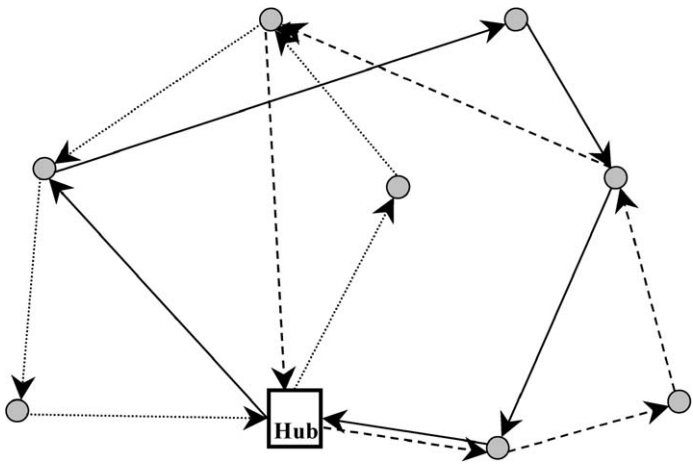


Fig. 3. Liner network design for a hub and spoke system. Example of three overlapping routes.

($D_{\mathrm{L}im}$) is the incremental demand for unloading (loading) in port i when the number of calls at that port increases from $m-1$ to m.

In the mathematical formulation, we use the following types of variables: the integer variable s represents the number of ships in operation and u_r, $r \in \mathcal{R}$, represents the number of voyages along route r during the planning horizon. The number of containers unloaded and loaded in port i on route r during the planning horizon is given by $q_{\mathrm{U}ir}$ and $q_{\mathrm{L}ir}$, $r \in \mathcal{R}$, $i \in \mathcal{N}_r$, respectively. The integer number of calls at port i is h_i, $i \in \mathcal{N}$, and finally, the binary variable g_{im}, $i \in \mathcal{N}$, $m \in \mathcal{M}$, is equal to 1 if port i is called at least m times during the planning horizon.

A liner network design model for a network with one hub and several spokes is as follows:

$$\max\Bigg[\Bigg(\sum_{r\in\mathcal{R}}\sum_{i\in\mathcal{N}_r}(R_{\mathrm{U}i}-C_{\mathrm{U}ir})q_{\mathrm{U}ir}\Bigg) + \Bigg(\sum_{r\in\mathcal{R}}\sum_{i\in\mathcal{N}_r}(R_{\mathrm{L}i}-C_{\mathrm{L}ir})q_{\mathrm{L}ir}\Bigg) - C_{\mathrm{F}}s - \sum_{r\in\mathcal{R}}C_{\mathrm{V}r}u_r\Bigg] \tag{3.31}$$

subject to

$$\sum_{r\in\mathcal{R}} T_{\mathrm{V}r}u_r - Ss \leqslant 0, \tag{3.32}$$

$$\sum_{i\in\mathcal{N}_r} q_{\mathrm{U}ir} - Qu_r \leqslant 0, \quad \forall r \in \mathcal{R}, \tag{3.33}$$

$$\sum_{j\in\mathcal{N}_{ir}^-} q_{\mathrm{L}jr} + \sum_{j\in\mathcal{N}_{ir}^+} q_{\mathrm{U}jr} - Qu_r \leqslant 0, \quad \forall r \in \mathcal{R}, i \in \mathcal{N}_r, \tag{3.34}$$

$$\sum_{r\in\mathcal{R}_i} u_r - h_i = 0, \quad \forall i \in \mathcal{N}, \tag{3.35}$$

$$\sum_{m\in\mathcal{M}} g_{im} - h_i = 0, \quad \forall i \in \mathcal{N}, \tag{3.36}$$

$$g_{i(m-1)} - g_{im} \geqslant 0, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}, \tag{3.37}$$

$$\sum_{r\in\mathcal{R}_i} q_{\mathrm{U}ir} - \sum_{m\in\mathcal{M}} D_{\mathrm{U}im}g_{im} \leqslant 0, \quad \forall i \in \mathcal{N}, \tag{3.38}$$

$$\sum_{r\in\mathcal{R}_i} q_{\mathrm{L}ir} - \sum_{m\in\mathcal{M}} D_{\mathrm{L}im}g_{im} \leqslant 0, \quad \forall i \in \mathcal{N}, \tag{3.39}$$

$$q_{\mathrm{U}ir}, q_{\mathrm{L}ir} \geqslant 0, \quad \forall r \in \mathcal{R}, i \in \mathcal{N}_r, \tag{3.40}$$

$$h_i, s, u_r \geqslant 0 \text{ and integer}, \quad \forall r \in \mathcal{R}, i \in \mathcal{N}, \tag{3.41}$$

$$g_{im} \in \{0,1\}, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}. \tag{3.42}$$

The objective function (3.31) maximizes the net revenue over the planning horizon. We calculate the number of needed ships in (3.32) in a way that might

be too simple. The constraints ensure that the total available sailing time for the total fleet of ships is larger than the sum of the voyages' times. We have not verified that the available time of the ships can be split in such a manner that each ship can perform an integer number of voyages during the planning horizon. Constraints (3.33) and (3.34) take care of the capacity when the ships leave the hub and the spokes on the route. Constraints (3.35) and (3.36) use the number of voyages along the routes to calculate the number of calls at each port. The precedence constraints (3.37) for the g_{im} variables are not needed if the incremental increase in the demand diminishes with increasing number of calls. The numbers of containers unloaded and loaded in the ports are bounded by the demand constraints (3.38) and (3.39). Finally, the formulation involves binary, integer and nonnegativity requirements on the variables in (3.40)–(3.42).

Bendall and Stent (2001) presented this model using a different notation and equal costs for loading and unloading containers. Their paper does not provide any information regarding how the model is solved. From the size of their practical example and the lack of information about the solution method, we conclude that they used some standard software for integer programming. After solving the stated model, they use heuristic methods to find a schedule for each ship. They report results for an application with Singapore as the hub and 6 spokes in East-Asia. The routes are different from the impression that the mathematical model gives, because they had 6 single spoke routes, one for each spoke and 2 routes with 2 spokes each. The demand data was for one week and it was assumed that the transportation pattern would be replicated for many weeks.

If we cannot guarantee that the incremental demand diminishes with increasing number of visits, then (3.35)–(3.39) can be reformulated in the following way. Some of the symbols will be redefined to avoid defining too many new ones. Now, let $D_{\mathrm{U}im}$ ($D_{\mathrm{L}im}$) be the unloading (loading) demand in port i when the number of calls in port i is m, and g_{im} is equal to 1 if port i is called exactly m times during the planning horizon.

These changes result in the following new or revised constraints:

$$\sum_{m\in\mathcal{M}} m g_{im} - \sum_{r\in\mathcal{R}_i} u_r = 0, \quad \forall i \in \mathcal{N}, \tag{3.43}$$

$$\sum_{m\in\mathcal{M}} g_{im} = 1, \quad \forall i \in \mathcal{N}, \tag{3.44}$$

$$\sum_{r\in\mathcal{R}_i} q_{\mathrm{U}ir} - \sum_{m\in\mathcal{M}} D_{\mathrm{U}im} g_{im} \leqslant 0, \quad \forall i \in \mathcal{N}, \tag{3.45}$$

$$\sum_{r\in\mathcal{R}_i} q_{\mathrm{L}ir} - \sum_{m\in\mathcal{M}} D_{\mathrm{L}im} g_{im} \leqslant 0, \quad \forall i \in \mathcal{N}. \tag{3.46}$$

Here (3.43) has replaced (3.35) and (3.36) and (3.44) is used instead of (3.37). After changing the meaning of the symbols, the last two constraints

above, (3.45) and (3.46), are unchanged from the original formulation. This reformulation might be useful when branching on g_{im} for one value of i and all values of m as one entity. Some solvers include this possibility, and this set of variables is then defined as a special ordered set of type one (SOS1 or S1). For a definition of such sets, see Chapter 9.3 in Williams (1999). For such sets some solvers will do binary branching by setting some of the variables equal to zero in one branch and setting the other variables equal to zero in the other branch. Such branching often results in a more evenly balanced branching tree. This in turn usually results in fewer branches to investigate.

3.3.3 Shuttle services

Ferries are often used to provide a shuttle service between a pair of ports. The ferries are often custom built to serve a particular route, fitting comfortably into available berths. Ferries may carry passengers, and usually can carry cars or trucks that are driven on and off board. Larger ferries that are designed to carry trucks or cars are called roll-on roll-off vessels. Very little research has been devoted to this area. A simulation model for ferry traffic among the Aegean Islands is described by Darzentas and Spyrou (1996). The model is used for decision support on a "what if" basis for regional development. By using the simulation model, they were able to evaluate the appropriateness of existing ferry routes, as well as new transportation scenarios, including the use of new technology vessels and changes in port capacities.

3.4 Design of maritime transport systems

In a maritime transport system, sea transport constitutes at least one vital link. An important strategic planning issue is the design of such systems. In the literature such systems are also referred to as maritime logistics systems or maritime supply chains. Reported research in the literature on such systems is scarce. We shall briefly discuss here one optimization-based application and a couple of simulation studies.

A real strategic and tactical industrial ocean-cargo shipping problem was studied by Mehrez et al. (1995). The problem involves the shipping of dry bulk products from a source port to transshipment ports, and then distribution of the products from the transshipment ports to the customers over land. The decisions made include the number and size of ships to charter in each time period during the planning horizon, the number and location of transshipment ports to use, and transportation routes from the transshipment ports to the customers. The problem is modeled and solved using a MIP model. Recommendations from this study were implemented by the client company.

Richetta and Larson (1997) present a problem regarding the design of New York City's refuse marine transport system. Waste trucks unload their cargo at land-based stations where refuse is placed into barges that are towed by tugboats to the Fresh Kills Landfill on Staten Island. They developed a discrete

event simulation model incorporating a complex dispatching module for decision support in fleet sizing and operational planning. This work is an extension of an earlier study by Larson (1988).

Another simulation study regarding maritime supply chain design can be found in Fagerholt and Rygh (2002). There, the problem is to design a seaborne system for transporting freshwater from Turkey to Jordan. The fresh water was to be transported by sea from Turkey to discharging buoy(s) off the coast of Israel, then in pipeline(s) to a tank terminal ashore and finally through a pipeline from Israel to Jordan. The study aimed at answering questions regarding the required number, capacity and speed of vessels, capacity and number of discharging buoys and pipelines, and the necessary capacity of the tank terminal.

Sigurd et al. (2005) discuss a problem where a group of companies, that need transport between locations on the Norwegian coastline and between Norway and The European Union, is focusing on reducing costs and decreasing transport lead-time by combining their shipments on the same ships. The companies need to analyze if there is a realistic possibility to switch some of their demand for transportation from road to sea. New transport solutions would need faster ships in order to substantially decrease the existing travel time. The underlying planning problem consists of finding recurring liner routes. These routes need to fit both with the quantity and frequency demanded by the companies.

3.5 Contract evaluation

This section discusses another important strategic problem faced by most shipping companies, namely contract evaluation. This problem is to some extent related to the fleet size and composition issue, and it consists of deciding whether to accept a specified long-term contract or not. The characteristics of this problem differ between tramp and liner operations, and this problem is of little relevance in an industrial operation.

For a *tramp shipping* company the problem is to decide whether to accept a Contract of Affreightment (a contract to carry specified quantities of cargo between specified ports within a specific time frame for an agreed payment per ton). In this case, the shipping company has to evaluate whether it has sufficient fleet tonnage to fulfill the contract commitments together with its existing commitments, and if so, whether the contract is profitable. To check if a contract will be profitable one also has to make assumptions about how the future spot market will develop for the given contract period. Typically, if a shipping company anticipates low spot rates, it will prefer to have as large contract coverage as possible or 'go short of tonnage' and vice versa. The authors are not aware of any published work in this area.

In the *liner shipping* industry these problems look slightly different. It is common that shippers buy a certain capacity for a given trade route. For instance in container freight transportation, which constitutes most of the liner shipping trade, it is not unusual that some of the bigger ocean carriers do between 80% and 95% of their business under such contracts. Most contracts between ocean

carriers and shippers are negotiated once a year, typically one or two months before the peak season of the major trade covered by the contract. A key parameter of a contract is the set of prices for the different cargoes between any pair of ports. The United States Ocean Shipping Reform Act of 1998 for the first time allows ocean carriers moving freight into and out of the US to enter into confidential contracts with shippers, and to charge different shippers different prices. This makes the problem of how to structure these prices relevant. This problem has many similarities with yield management in the airline industry. Kleywegt (2003) presents a model that can be used to support such decisions before and during contract negotiations. A somewhat similar problem can be found for cruise lines. Ladany and Arbel (1991) present four models for determining the optimal price differentiation strategy that a cruise liner should follow in order to maximize its profit for four different situations. A price differentiation strategy means that customers belonging to different market segments would pay different prices for identical cabins. Also this problem is similar to yield management in airlines.

4 Tactical planning in maritime transportation

At the tactical planning level we concentrate on medium-term decisions, and the focus of this level in maritime transportation is on routing and scheduling. Therefore, most of this section is devoted to these planning issues. We start this section by presenting some classical industrial and tramp ship scheduling problems and give arc flow formulations of these problems in Section 4.1. Then in Section 4.2 we discuss frequently used solution methods for solving ship routing and scheduling problems. Throughout the presentation of problems, formulations and solution approaches we refer to important research in industrial and tramp ship scheduling, as we deem appropriate. In Section 4.3 we present several tactical planning problems and applications in maritime supply chains, where sea transport constitutes at least one vital part of the supply chain. Fleet deployment in liner shipping is presented and discussed in Section 4.4, whereas barge scheduling on inland waterways is presented in Section 4.5. Section 4.6 is dedicated to naval vessel scheduling, while in Section 4.7 we briefly discuss ship management.

4.1 Scheduling problems for industrial and tramp shipping

As described in Section 2, in industrial shipping the cargo owner or shipper controls the ships. Industrial operators try to ship all their cargoes at minimum cost. Tramp ships follow the available cargoes like a taxi. A tramp shipping company may have a certain amount of contract cargoes that it is committed to carry, and tries to maximize the profit from optional cargoes. From an OR point of view the structure of the planning challenges for these two modes of operation is very similar regarding the underlying mathematical models and

solution approaches. Therefore we treat these modes of operations together in this section. During the last decades there has been a shift from industrial to tramp shipping (see Christiansen et al., 2004 and Section 7). In Section 7 we discuss some reasons for the shift from industrial to tramp shipping. Perhaps the main reason is that many cargo owners are now focusing on their core business and have outsourced other activities like transportation to independent shipping companies. From the shipper's perspective, this outsourcing has resulted in reduced risk. Most contributions in the OR literature are for industrial shipping, while only a few are in the tramp sector. The main reason for the minimal attention to tramp scheduling in the literature may be that historically the tramp market was operated by a large number of small operators, even though this is not the case anymore.

In this section we present classes of real ship routing and scheduling problems. We start with the simplest type of problems in Section 4.1.1 dealing with routing and scheduling of full shiploads. Here just one cargo is onboard the ship at a time. We extend this problem to multiple cargoes onboard at the same time, where each of the cargoes has a fixed size. This problem is addressed in Section 4.1.2. We continue in Section 4.1.3 with similar problems but where flexible cargo sizes are allowed. In Section 4.1.4 we present routing and scheduling problems where multiple nonmixable products can be carried simultaneously, and the ship capacity is split into separate compartments. Typical tramp shipping characteristics concerning contracted and optional cargoes are considered in Section 4.1.5. Finally, we discuss the use of spot charters in Section 4.1.6.

In practice, at the beginning of the planning horizon the ships in the fleet may be occupied with prior tasks. For all the classes of problems described in this section we find the first point in time where the ship is available for loading a new cargo during the planning horizon, and we assume that at that time the ship is empty.

4.1.1 Full shiploads

In some market segments, the ship is loaded to its capacity in a loading port and the cargo is transported directly to its unloading port. A typical example is the transportation of crude oil.

The objective of an industrial ship scheduling problem for full shipload cargoes is to minimize the sum of the costs for all the ships in the fleet while ensuring that all cargoes are lifted from their loading ports to their corresponding unloading ports. Time windows are usually imposed for both loading and unloading the cargoes.

In such an operation, an industrial shipping company usually operates a heterogeneous fleet of ships with specific ship characteristics including different cost structures and load capacities. In the short-term, it is impractical to change the fleet size. Therefore, we are concerned with the operations of a given number of ships within the planning horizon. The fixed cost of the fleet can be disregarded as it has no influence on the planning of optimal routes and sched-

ules. We consider the case where the fleet has sufficient capacity to serve all committed cargoes during the planning horizon. The ships are charged port and channel tolls when visiting ports and passing channels, and these costs depend on the size of the ship. The remaining variable sailing costs consist mainly of fuel and oil costs, and depend usually on the ship size.

The quantity of a particular cargo is given and the corresponding loading and unloading port of that cargo are known, so the time from arrival at the loading port until the time of departure from the unloading port can be easily calculated.

In the case where a ship can carry only one cargo at a time but the ship is not necessary filled up each time, the underlying planning problem is identical to the problem of full shiploads.

Example 4.1. Consider the following simplified example of a route from a solution to a full shipload planning problem. In this planning problem several ships are going to service a set of cargoes. In the optimal solution, one ship is going to lift cargoes 1, 2, and 3. In Table 5, information about the loading and unloading ports is given for each of the cargoes. In addition, we specify the quantity of each of the cargoes. Notice that not all cargo sizes are equal to the capacity of the ship. Two of the cargoes have a quantity equal to half the capacity of the ship. In reality, the utilization of the ship is too low, but this case is a basis for another problem presented later on in this section. For the sake of simplicity, the time windows information is omitted in this example.

The geographical picture of the ports is given in Figure 4(a), while the physical planned route for the ship is shown in Figure 4(b). The physical planned route is the shortest route for this set of cargoes. Notice that the sequence of cargoes in the optimal solution might be different when we consider the time windows. Finally, in Figure 4(c), we see the load onboard the ship at departure from the respective ports for the planned route.

In the mathematical description of the problem, let $\mathcal{N}$ be the set of cargoes indexed by i. Cargo i is represented by a node in a network, and this node includes one loading port and one unloading port for cargo i. Further, let $\mathcal{V}$ be the set of ships in the fleet indexed by v. The set $(\mathcal{N}_v, \mathcal{A}_v)$ is the network associated with a specific ship v, where $\mathcal{N}_v$ and $\mathcal{A}_v$ represent the sets of the nodes and arcs, respectively. Not all ships can visit all ports and take all cargoes, and $\mathcal{N}_v = \{\text{feasible nodes for ship } v\} \cup \{o(v), d(v)\}$. Here, $o(v)$ and $d(v)$ are an artificial origin cargo and an artificial destination cargo for ship v, respectively. If the ship is not used, $d(v)$ will be serviced just after $o(v)$. The set $\mathcal{A}_v$ contains all feasible arcs for ship v, which is a subset of $\{i \in \mathcal{N}_v\} \times \{i \in \mathcal{N}_v\}$. This set will be calculated based on time constraints and other restrictions. The arc (i, j) connects cargo i and cargo j, where cargo i will be serviced just before cargo j if the arc is used.

Let us look again at Example 4.1. Figure 5 shows the route of this example (marked with bold lines) drawn over the underlying network. The ship leaves

Table 5.
Cargo information for Examples 4.1 and 4.2

	Loading port	Unloading port	Quantity
Cargo 1	A	C	$\frac{1}{2}$ ship
Cargo 2	D	E	full ship
Cargo 3	B	D	$\frac{1}{2}$ ship

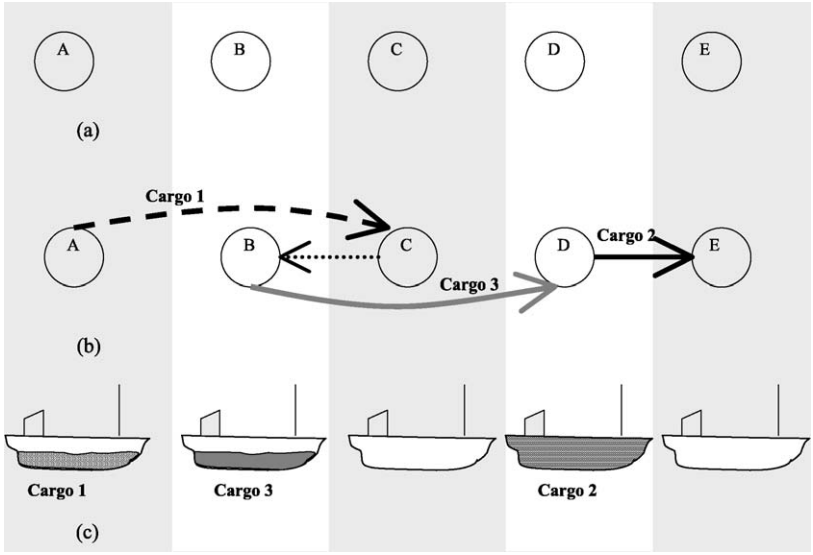


Fig. 4. (a) Geographical picture of the ports for Examples 4.1 and 4.2. (b) Physical route for the ship for Example 4.1. (c) Load onboard the ship at departure for Example 4.1.

the artificial origin cargo node in the beginning of its route and lifts cargo 1 that is represented by node Cargo 1. The route is then followed by node Cargo 3, node Cargo 2, and finally the artificial destination cargo node. The other arcs are possible precedence combinations between the cargoes given in this example.

For each arc, T_{Sijv} represents the calculated time for ship v from the arrival at the loading port for cargo i until the arrival at the loading port for cargo j. It includes the sum of the time for loading and unloading cargo i, the sailing time between ports related to cargo i and the sailing time from the unloading port for cargo i to the loading port for cargo j. Let $[T_{MNiv}, T_{MXiv}]$ denote the time window for ship v associated with the loading port for cargo i, where T_{MNiv} is the earliest time for start of service, while T_{MXiv} is the latest time. In the

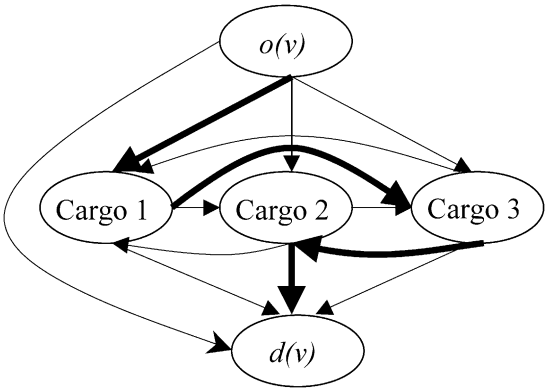


Fig. 5. The route of Example 4.1 drawn over the underlying network.

underlying real problem these data are seldom specified for each ship v but are appropriate in the mathematical model due to a preprocessing phase. The variable sailing and port costs are represented by C_{ijv}.

In the mathematical formulation, we use the following types of variables: the binary flow variable x_{ijv}, $v \in \mathcal{V}$, $(i, j) \in \mathcal{A}_v$, equals 1, if ship v services cargo i just before cargo j, and 0 otherwise. This flow variable determines which ship takes a particular cargo. The time variable t_{iv}, $v \in \mathcal{V}$, $i \in \mathcal{N}_v$, represents the time at which service begins at the loading port of cargo i with ship v.

The arc flow formulation of the industrial ship scheduling problem with full shiploads is as follows:

$$\min \sum_{v \in \mathcal{V}} \sum_{(i,j) \in \mathcal{A}_v} C_{ijv} x_{ijv} \tag{4.1}$$

subject to

$$\sum_{v \in \mathcal{V}} \sum_{j \in \mathcal{N}_v} x_{ijv} = 1, \quad \forall i \in \mathcal{N}, \tag{4.2}$$

$$\sum_{j \in \mathcal{N}_v} x_{o(v)jv} = 1, \quad \forall v \in \mathcal{V}, \tag{4.3}$$

$$\sum_{i \in \mathcal{N}_v} x_{ijv} - \sum_{i \in \mathcal{N}_v} x_{jiv} = 0, \quad \forall v \in \mathcal{V}, j \in \mathcal{N}_v \backslash \{o(v), d(v)\}, \tag{4.4}$$

$$\sum_{i \in \mathcal{N}_v} x_{id(v)v} = 1, \quad \forall v \in \mathcal{V}, \tag{4.5}$$

$$x_{ijv}(t_{iv} + T_{\mathrm{S}ijv} - t_{jv}) \leqslant 0, \quad \forall v \in \mathcal{V}, (i, j) \in \mathcal{A}_v, \tag{4.6}$$

$$T_{\mathrm{MN}iv} \leqslant t_{iv} \leqslant T_{\mathrm{MX}iv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_v, \tag{4.7}$$

$$x_{ijv} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, (i, j) \in \mathcal{A}_v. \tag{4.8}$$

The objective function (4.1) minimizes the costs of operating the fleet. Constraints (4.2) ensure that all cargoes that the shipping company has committed itself to carry are serviced. Constraints (4.3)–(4.5) describe the flow on the sailing route used by ship v. Constraints (4.3) and (4.5) ensure that ship v services the artificial origin cargo and the artificial destination cargo once, respectively. Constraints (4.6) describe the compatibility between routes and schedules. The time for start of service of cargo j cannot be less than the sum of the start time of cargo i and the service time for loading, transporting and unloading cargo i and the sailing time from the unloading port for cargo i to the loading port for cargo j with ship v, if ship v is really servicing cargo i just before cargo j. Constraints (4.6) contain an inequality sign because waiting time is permitted before the start of service in a port. The time window constraints are given by constraints (4.7). For the artificial origin cargo, this time window is collapsed to the value when ship v is available for new cargoe(s) during the planning horizon. If ship v is not servicing cargo i, we get an artificial starting time within the time windows for that (i, v)-combination. This means that we get a starting time for each (i, v)-combination. However, just the starting time associated with ship v actually lifting the particular cargo i is real. Finally, the formulation involves binary requirements (4.8) on the flow variables.

This industrial ship scheduling problem for full shipload cargoes corresponds to a multitraveling salesman problem with time windows (see Desrosiers et al., 1995).

The model (4.1)–(4.8) is still valid if the planning problem involves cargoes that are not equal to the capacity of the ship but a ship can carry only one cargo at a time. The set $\mathcal{N}_v$ gives the cargoes that can be serviced by ship v. For this variant of the problem, the set $\mathcal{N}_v$ is calculated based on the capacity of the ship and the load quantity of cargo i.

The quantities of some cargoes might be given in an interval, and the cargo size is then determined by the ship capacity a priori for each cargo and ship combination. Relative revenues for loading larger cargo quantities for a cargo i due to larger ship capacity can be included in C_{ijv}.

The load of the ship might in some cases be first loaded in several loading ports in the same region and unloaded in one or several ports. The model (4.1)–(4.8) is also valid for such a situation. However, the calculated sailing times have to be adjusted such that times in all ports are included. Now, the time variable t_{iv} represents the time at which service begins at the *first* loading port for cargo i with ship v.

In the literature, we find several studies on the industrial ship scheduling problems with full shipload cargoes. Brown et al. (1987) describe such a problem where a major oil company is shipping crude oil from the Middle East to Europe and North America. Fisher and Rosenwein (1989) study a problem that is conceptually quite similar to the one in Brown et al. (1987). Here, a fleet of ships controlled by the Military Sealift Command of the US Navy is engaged in pickup and delivery of various bulk cargoes. Each cargo may have up to three loading points which are often the same port or nearby ports and

up to three unloading points that are frequently close to each other. In contrast to Brown et al. (1987), each cargo may not be a full shipload. However, at most one cargo is on a vessel at any time. Therefore, the same model is still valid. Another similar problem of shipping crude oil is studied by Perakis and Bremer (1992).

4.1.2 Multiple cargoes with fixed cargo size

Here we present an industrial ship routing and scheduling problem where several cargoes are allowed to be onboard the ship at the same time. The objective of the scheduling problem is to minimize the sum of the costs for all the ships in the fleet while ensuring that all cargoes are lifted from their loading ports to their corresponding unloading ports. Each cargo consists of a designated number of units of a product or a commodity. Time windows are normally imposed for both the pickup and delivery of the cargoes. The ship capacities, the cargo type and quantities are such that the ships may carry several cargoes simultaneously. This means that another loading port can be visited with some cargoes still onboard. We assume that the cargoes are compatible with each other.

Example 4.2. This example is based on Example 4.1. We have the same cargo information as given in Table 5, and the geographical picture of the ports is shown in Figure 4(a). However, multiple cargoes can be carried simultaneously. Figure 6(a) shows the physical route for the ship.

Cargo 1 is lifted in port A and the ship sails to port B to load Cargo 3. On departure the ship is fully loaded with two cargoes. Figure 6(b) shows the load onboard the ship upon departure from each port.

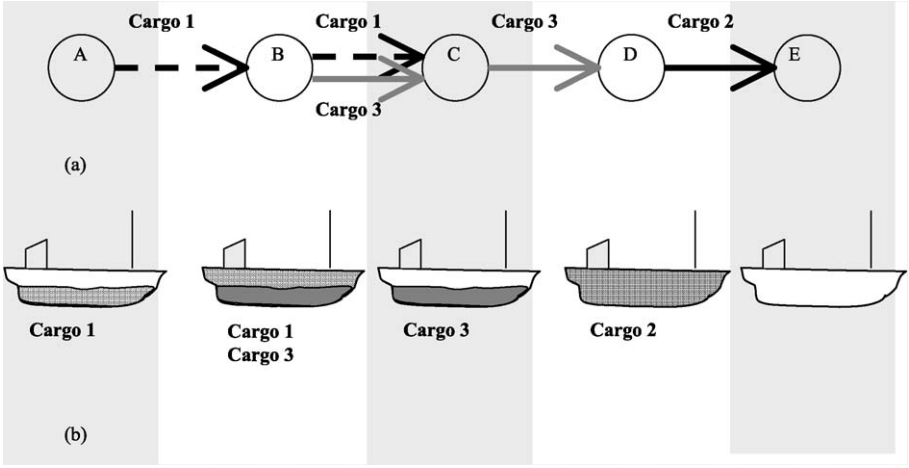


Fig. 6. (a) Physical route for a ship with multiple cargoes onboard for Example 4.2. (b) Load onboard the ship at departure for Example 4.2.

We have the same conditions for the fleet as for the problem described in Section 4.1.1, concerning a heterogeneous fixed fleet with various variable costs. In addition, we assume that the sailing costs do not depend on the load onboard the ship.

In the mathematical description of the problem also here each cargo is represented by an index i. However, associated with the loading port of cargo i, there is a node i, and with the corresponding unloading port a node $N + i$, where N is the number of cargoes that has to be serviced during the planning horizon. Note that different nodes may correspond to the same physical port. Let $\mathcal{N}_{\mathrm{P}} = \{1, \ldots, N\}$ be the set of loading (or pickup) nodes and $\mathcal{N}_{\mathrm{D}} = \{N + 1, \ldots, 2N\}$ be the set of unloading (or delivery) nodes, and define $\mathcal{N} = \mathcal{N}_{\mathrm{P}} \cup \mathcal{N}_{\mathrm{D}}$. $\mathcal{V}$ is the set of ships in the fleet indexed by v. Then $(\mathcal{N}_v, \mathcal{A}_v)$ is the network associated with a specific ship v. Here, $\mathcal{N}_v = \{\text{feasible nodes for ship } v\} \cup \{o(v), d(v)\}$ is the set of ports that can be visited by ship v and $o(v)$ and $d(v)$ are the artificial origin depot and artificial destination depot of ship v, respectively. Geographically, the artificial origin depot $o(v)$ can be either a port or a point at sea, while the artificial destination depot $d(v)$ is the last planned unloading port for ship v. If the ship is not used $d(v)$ will represent the same location as $o(v)$. Here $\mathcal{A}_v$ contains the set of all feasible arcs for ship v, which is a subset of $\{i \in \mathcal{N}_v\} \times \{i \in \mathcal{N}_v\}$. This set will be calculated based on capacity and time constraints, and other restrictions such as those based on precedence of loading and unloading nodes for the same cargo. From these calculations, we can extract the sets $\mathcal{N}_{\mathrm{P}v} = \mathcal{N}_{\mathrm{P}} \cap \mathcal{N}_v$ and $\mathcal{N}_{\mathrm{D}v} = \mathcal{N}_{\mathrm{D}} \cap \mathcal{N}_v$ consisting of loading and unloading nodes that ship v may visit, respectively.

Let us refer back to Example 4.2. In the underlying network for the example, we introduce two nodes for each of the cargoes. This means that Cargo 1 is represented by the loading node 1 and the unloading node $N + 1$. The loading port for Cargo 2 and the unloading port for Cargo 3 are the same physical port. That means that both node 2 and node $N + 3$ represent port D. Figure 7 shows the route of this example (marked with bold lines). The other arcs are left out of the figure for sake of clarity. In general, there will be arcs from $o(v)$ to all loading ports and $d(v)$. In addition, we will have arcs into $d(v)$ from $o(v)$ and all unloading ports. The network for the real loading and unloading ports will be complete except for arcs from each of the unloading ports $N + i$ to the corresponding loading port i. The sequence of nodes for this example is as follows: $o(v)$–1–3–$(N + 1)$–$(N + 3)$–2–$(N + 2)$–$d(v)$.

The fixed cargo quantity for cargo i is given by Q_i, while the capacity of ship v is given by $V_{\mathrm{CAP}v}$. For each arc, $T_{\mathrm{S}ijv}$ represents the sum of the calculated sailing time from node i to node j with ship v and the service time at node i. Let $[T_{\mathrm{MN}iv}, T_{\mathrm{MX}iv}]$ denote the time window associated with node i and ship v. The variable sailing and port costs are represented by C_{ijv}.

In the mathematical formulation, we use the following types of variables: the binary flow variable x_{ijv}, $v \in \mathcal{V}$, $(i, j) \in \mathcal{A}_v$, equals 1, if ship v sails from node i directly to node j, and 0 otherwise. The time variable t_{iv}, $v \in \mathcal{V}$, $i \in \mathcal{N}_v$,

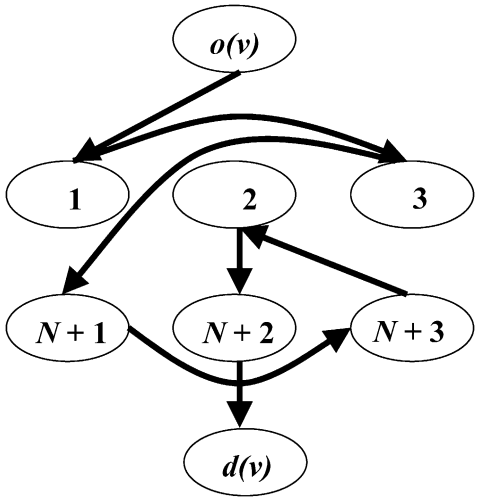


Fig. 7. The route of Example 4.2.

represents the time at which service begins at node i, while variable l_{iv}, $v \in \mathcal{V}$, $i \in \mathcal{N}_v \backslash \{d(v)\}$, gives the total load onboard ship v just after the service is completed at node i.

The arc flow formulation of the industrial ship scheduling problem with fixed cargo sizes is as follows:

$$\min \sum_{v \in \mathcal{V}} \sum_{(i,j) \in \mathcal{A}_v} C_{ijv} x_{ijv} \tag{4.9}$$

subject to

$$\sum_{v \in \mathcal{V}} \sum_{j \in \mathcal{N}_v} x_{ijv} = 1, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, \tag{4.10}$$

$$\sum_{j \in \mathcal{N}_{\mathrm{P}v} \cup \{d(v)\}} x_{o(v)jv} = 1, \quad \forall v \in \mathcal{V}, \tag{4.11}$$

$$\sum_{i \in \mathcal{N}_v} x_{ijv} - \sum_{i \in \mathcal{N}_v} x_{jiv} = 0,$$

$$\forall v \in \mathcal{V}, j \in \mathcal{N}_v \backslash \{o(v), d(v)\}, \tag{4.12}$$

$$\sum_{i \in \mathcal{N}_{\mathrm{D}v} \cup \{o(v)\}} x_{id(v)v} = 1, \quad \forall v \in \mathcal{V}, \tag{4.13}$$

$$x_{ijv}(t_{iv} + T_{\mathrm{S}ijv} - t_{jv}) \leqslant 0, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v, \tag{4.14}$$

$$T_{\mathrm{MN}iv} \leqslant t_{iv} \leqslant T_{\mathrm{MX}iv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_v, \tag{4.15}$$

$$x_{ijv}(l_{iv} + Q_j - l_{jv}) = 0, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}, \tag{4.16}$$

$$x_{i,N+j,v}(l_{iv} - Q_j - l_{N+j,v}) = 0,$$
$$\forall v \in \mathcal{V}, (i, N+j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}, \tag{4.17}$$
$$l_{o(v)v} = 0, \quad \forall v \in \mathcal{V}, \tag{4.18}$$
$$\sum_{j \in \mathcal{N}_v} Q_i x_{ijv} \leqslant l_{iv} \leqslant \sum_{j \in \mathcal{N}_v} V_{\mathrm{CAP}v} x_{ijv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.19}$$
$$0 \leqslant l_{N+i,v} \leqslant \sum_{j \in \mathcal{N}_v} (V_{\mathrm{CAP}v} - Q_i) x_{N+i,jv},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.20}$$
$$t_{iv} + T_{\mathrm{S}i,N+i,v} - t_{N+i,v} \leqslant 0, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.21}$$
$$\sum_{j \in \mathcal{N}_v} x_{ijv} - \sum_{j \in \mathcal{N}_v} x_{j,N+i,v} = 0, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.22}$$
$$x_{ijv} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, (i, j) \in \mathcal{A}_v. \tag{4.23}$$

The objective function (4.9) minimizes the costs of operating the fleet. Constraints (4.10) ensure that all cargoes that the shipping company has committed itself to carry are serviced. Constraints (4.11)–(4.13) describe the flow on the sailing route used by ship v. Constraints (4.14) describe the compatibility between routes and schedules. The starting time of the service at node j cannot be less than the sum of the starting time and the loading time at node i and the sailing time from i to j with ship v, if ship v is really sailing between these two nodes. The time window constraints are given by (4.15). If ship v is not visiting node i, we will get an artificial starting time within the time windows for that (i, v)-combination. Introduction of artificial starting times is practical, due to constraints (4.21). Constraints (4.16) and (4.17) give the relationship between the binary flow variables and the ship load at each loading and unloading port, respectively. The initial load condition for each ship is given by (4.18). The ship is empty at the beginning of the planning horizon as mentioned in the opening of Section 4.1. Constraints (4.19) and (4.20) represent the ship capacity intervals at loading and unloading nodes, respectively. Constraints (4.20) can be omitted from the model since the upper bound can never be exceeded due to constraints (4.19) and the precedence and coupling constraints (4.21) and (4.22). The precedence constraints forcing node i to be visited before node $N+i$ are given in (4.21). For both constraints (4.14) and (4.21), the constraints appear only if the beginning of the time window for nodes j and $N+i$, respectively, is less than the earliest calculated arrival time at the node. Along with the coupling constraints (4.22), constraints (4.21) ensure that the same ship v visits both node i and $N+i$, $i \in \mathcal{N}_{\mathrm{P}v}$. Finally, the formulation involves binary requirements (4.23) on the flow variables.

We find a few applications for this industrial shipping problem with fixed cargo quantities in the literature. Fagerholt and Christiansen (2000a) study a multiproduct scheduling problem. They extend the model presented here, and include allocation of cargoes to different flexible cargo holds. For more details,

see Section 4.1.4. Further, Christiansen and Fagerholt (2002) present a real ship scheduling problem which is based on the model (4.9)–(4.23). In addition, they focus on two important issues in the shipping industry, namely ports closed at night and over weekends and long loading or unloading operations. This study is described in more detail in Section 6.

The multiple cargo with fixed cargo size ship scheduling problem is also studied by Psaraftis (1988) for the US Military Sealift Command. The objective is to allocate cargo ships to cargoes so as to ensure that all cargoes arrive at their destinations as planned. Constraints that have to be satisfied include loading and unloading time windows for the cargoes, ship capacity and cargo/ship/port compatibility. The problem is dynamic, because in a military mobilization situation anything can change in real time. The paper focuses on the dynamic aspects of the problem and the algorithm that is developed is based on the "rolling horizon" approach. Later, Thompson and Psaraftis (1993) applied a new class of neighborhood search algorithms to a variety of problems, including the problem of the US Military Sealift Command.

4.1.3 Multiple cargoes with flexible cargo size

For many real ship scheduling problems, the cargo quantity is given in an interval and the shipping company can choose the actual load quantity that best fits its fleet and schedule. For such problems, the minimum cost problem is transferred to a maximum profit problem. Apart from these issues, the problem is identical to the problem described in Section 4.1.2. We use the same mathematical notation and the same type of network representation as in Figure 7. However, we need the following additional notation:

The variable quantity interval is given by $[Q_{\mathrm{MN}i}, Q_{\mathrm{MX}i}]$, where $Q_{\mathrm{MN}i}$ is the minimum quantity to be lifted, while $Q_{\mathrm{MX}i}$ is the maximum quantity for cargo i. The time required to load or unload one unit of a cargo at node i is given by $T_{\mathrm{Q}i}$. The node can either be a loading or unloading node, which means that the time per unit might be different for loading and unloading. Here $T_{\mathrm{S}ijv}$ is just the sailing time between the two ports and does not include the service time in any of the ports.

We need an additional continuous variable q_{iv}, $v \in \mathcal{V}$, $i \in \mathcal{N}_{\mathrm{P}v}$, that represents the quantity of cargo i, when cargo i is lifted by ship v and loaded at node i and unloaded at node $N+i$. The revenue of carrying a cargo is normally the cargo quantity q_{iv} multiplied by a revenue per unit of cargo P_i. However, in some cases the revenue from a cargo may be a lump sum or another function of the cargo quantity, and then the objective function becomes nonlinear. In the following mathematical formulation of the objective function we use a linear term for the revenue from carrying the cargoes.

The ship scheduling problem with flexible cargo sizes is formulated as follows:

$$\max\left[\sum_{v\in\mathcal{V}}\sum_{i\in\mathcal{N}_{\mathrm{P}v}} P_i q_{iv} - \sum_{v\in\mathcal{V}}\sum_{(i,j)\in\mathcal{A}_v} C_{ijv}x_{ijv}\right] \tag{4.24}$$

subject to

$$\sum_{v\in\mathcal{V}}\sum_{j\in\mathcal{N}_v} x_{ijv} = 1, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, \tag{4.25}$$

$$\sum_{j\in\mathcal{N}_{\mathrm{P}v}\cup\{d(v)\}} x_{o(v)jv} = 1, \quad \forall v \in \mathcal{V}, \tag{4.26}$$

$$\sum_{i\in\mathcal{N}_v} x_{ijv} - \sum_{i\in\mathcal{N}_v} x_{jiv} = 0, \quad \forall v \in \mathcal{V}, j \in \mathcal{N}_v\backslash\{o(v), d(v)\}, \tag{4.27}$$

$$\sum_{i\in\mathcal{N}_{\mathrm{D}v}\cup\{o(v)\}} x_{id(v)v} = 1, \quad \forall v \in \mathcal{V}, \tag{4.28}$$

$$x_{ijv}(t_{iv} + T_{\mathrm{Q}i}q_{iv} + T_{\mathrm{S}ijv} - t_{jv}) \leqslant 0, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v \mid i \in \mathcal{N}_{\mathrm{P}v} \cup o(v), \tag{4.29}$$

$$x_{N+i,jv}(t_{N+i,v} + T_{\mathrm{Q}N+i}q_{iv} + T_{\mathrm{S}N+i,jv} - t_{jv}) \leqslant 0, \quad \forall v \in \mathcal{V}, (N+i,j) \in \mathcal{A}_v \mid i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.30}$$

$$T_{\mathrm{MN}iv} \leqslant t_{iv} \leqslant T_{\mathrm{MX}iv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_v, \tag{4.31}$$

$$x_{ijv}(l_{iv} + q_{jv} - l_{jv}) = 0, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}, \tag{4.32}$$

$$x_{i,N+j,v}(l_{iv} - q_{jv} - l_{N+j,v}) = 0, \quad \forall v \in \mathcal{V}, (i,N+j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}, \tag{4.33}$$

$$\sum_{j\in\mathcal{N}_v} Q_{\mathrm{MN}i}x_{ijv} \leqslant q_{iv} \leqslant \sum_{j\in\mathcal{N}_v} Q_{\mathrm{MX}i}x_{ijv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.34}$$

$$l_{o(v)v} = 0, \quad \forall v \in \mathcal{V}, \tag{4.35}$$

$$q_{iv} \leqslant l_{iv} \leqslant \sum_{j\in\mathcal{N}_v} V_{\mathrm{CAP}v}x_{ijv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.36}$$

$$0 \leqslant l_{N+i,v} \leqslant \sum_{j\in\mathcal{N}_v} V_{\mathrm{CAP}v}x_{N+i,jv} - q_{iv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.37}$$

$$t_{iv} + T_{\mathrm{Q}i}q_{iv} + T_{\mathrm{S}i,N+i,v} - t_{N+i,v} \leqslant 0, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.38}$$

$$\sum_{j\in\mathcal{N}_v} x_{ijv} - \sum_{j\in\mathcal{N}_v} x_{j,N+i,v} = 0, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.39}$$

$$x_{ijv} \in \{0,1\}, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v. \tag{4.40}$$

The objective function (4.24) maximizes the profit gained by operating the fleet. The constraints (4.25)–(4.40) are equivalent to (4.10)–(4.23), apart from

the following constraints. The constraints ensuring feasible time schedules are split into constrains for loading in port i, (4.29), and unloading in port $N + i$, (4.30). These constraints are adjusted for the variable loading time at port i. Variable q_{iv} is not defined for $i = o(v)$, so the term $T_{Qi}q_{iv}$ does not exist for $i = o(v)$ in constraints (4.29). Here, constraints (4.32) and (4.33) include a variable load quantity instead of the fixed quantity in constraints (4.16) and (4.17). In constraints (4.34) the load quantity interval is defined for each cargo i. The load variable q_{iv} is forced to 0 by (4.34) if cargo i is not lifted by ship v. Constraints (4.36)–(4.38) are adjusted for the variable load quantity.

A ship scheduling problem with flexible cargo sizes is studied by Brønmo et al. (2007) for transportation of bulk cargoes by chemical tankers and has many similarities to the problem described here. The solution method is based on a set partitioning approach that gives optimal solutions to the problem. Korsvik et al. (2007) solve the same problem by using a multistart local search heuristic.

There are operations where a ship can carry only one cargo at a time, but the ship is not necessarily filled up each time and the cargo quantity is given in an interval. For this situation, we still have variable load quantities and arrival times as in the model of this section. However, we do not need nodes for both loading and unloading ports, but just a common node representing the cargo as we did in the model of Section 4.1.1.

4.1.4 Multiple products

In Sections 4.1.1–4.1.3 we assumed that the cargoes consist of mixable products that can be loaded onboard regardless of the type of product already onboard. In addition, different cargoes are compatible with each other. However, often multiple nonmixable products are carried onboard a ship simultaneously. In such cases the cargo carrying space of the vessel must be divided into separate tanks (compartments or holds) that are usually fixed. For example, a large chemical tanker may have from 20 to 50 tanks. We start with considering the case where the cargo tanks of the ship are of equal size. In reality, this is seldom the case. However, it may be possible to separate the tanks into sets that are of about equal size. If the ship has many tanks, this assumption is reasonable. In addition, we assume that the cargo consists of mixable products, but different cargoes have to be stored in different tanks.

In the mathematical description of the problem, we need the following notation: the number of tanks (or cargo holds) of ship v is given by H_v and the capacity of a tank (hold) of ship v is given by $H_{CAPv} = V_{CAPv}/H_v$. As the ship is assumed empty at the first time it is available for scheduling during the planning horizon, the number of tanks (holds) occupied is also 0. Variable h_{iv}, $v \in \mathcal{V}$, $i \in \mathcal{N}_v$, represents the number of tanks (holds) occupied after servicing node i by ship v. We still use the continuous variable q_{iv}, $v \in \mathcal{V}$, $i \in \mathcal{N}_{Pv}$, representing the quantity of cargo i, when cargo i is lifted by ship v and loaded at node i and unloaded at node $N + i$.

In order to allow several different nonmixable cargoes onboard simultaneously, we need the following constraints added to formulation (4.24)–(4.40):

$$x_{ijv}\left(h_{iv} + \left\lceil \frac{q_{jv}}{H_{\mathrm{CAP}v}} \right\rceil - h_{jv}\right) = 0, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}, \tag{4.41}$$

$$x_{i,N+j,v}\left(h_{iv} - \left\lceil \frac{q_{jv}}{H_{\mathrm{CAP}v}} \right\rceil - h_{N+j,v}\right) = 0, \quad \forall v \in \mathcal{V}, (i,N+j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}, \tag{4.42}$$

$$\sum_{j \in \mathcal{N}_v} \left\lceil \frac{q_{iv}}{H_{\mathrm{CAP}v}} \right\rceil x_{ijv} \leqslant h_{iv} \leqslant \sum_{j \in \mathcal{N}_v} H_v x_{ijv}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.43}$$

$$0 \leqslant h_{N+i,v} \leqslant \sum_{j \in \mathcal{N}_v} H_v x_{N+i,jv} - \left\lceil \frac{q_{iv}}{H_{\mathrm{CAP}v}} \right\rceil, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_{\mathrm{P}v}, \tag{4.44}$$

$$h_{o(v)v} = 0, \quad \forall v \in \mathcal{V}, \tag{4.45}$$

$$h_{iv} \in [0, H_v] \text{ and integer}, \quad \forall v \in \mathcal{V}, i \in \mathcal{N}_v. \tag{4.46}$$

Constraints (4.41) and (4.42) describe the compatibility between routes and the number of occupied tanks when the arrival node is a loading port and an unloading port, respectively. The intervals of the number of occupied tanks after servicing the loading and unloading nodes are given in constraints (4.43) and (4.44), respectively. Next, constraints (4.45) impose the initial tank occupancy condition for each ship. Finally, the integer requirements for the tank number variables are given. The integer interval $[0, H_v]$ in (4.46) can be reduced by information from (4.34) and (4.44).

For problems with multiple, nonmixable, products for a cargo, the allocation of products to the various tanks is normally needed. For transportation of liquid products, the quantity has to be flexible due to stability considerations and to prevent product sloshing in partially empty tanks.

In the literature, Scott (1995) presents a problem involving the shipping of refined oil products from a refinery to several depots. Several types of tankers/ships with fixed tanks enable different products to be carried on the same voyage (without mixing them). Another study with multiple products is given by Bausch et al. (1998). They present a decision support system for medium-term scheduling where a fleet of coastal tankers and barges are transporting liquid bulk products among plants, distribution centers, and industrial customers. A set of cargoes has to be conveyed by the available fleet of vessels and each cargo consists of an ordered volume of up to five products. The vessels may have up to seven fixed tanks, thus allowing a cargo consisting of several products to be lifted by the same ship. When multiple cargoes are carried simultaneously, different cargoes of the same product are stowed in different tanks. Such cargoes are not mixed in order to eliminate the need for measuring the unloaded quantity at the multiple unloading ports. A similar

problem is studied by Sherali et al. (1999) describing a ship scheduling problem where crude oil and a number of refined oil-related products are to be shipped from ports in Kuwait to customers around the world. Here, each cargo is a full shipload of a compartmentalized group of products, and is characterized by its mix (oil, refined products, etc.), loading port, loading date, unloading port, and unloading date. The ships have multiple tanks of different sizes, so they introduce a flow variable that is 1 if a particular tank carries a particular product on a particular leg (i, j) with ship v. The model is extended compared to the one presented here and includes the allocation of product quantities to tanks.

Recently, Jetlund and Karimi (2004) presented a similar problem for multicompartment tankers engaged in shipping bulk liquid chemicals. They present a mixed-integer linear programming formulation using variable-length time slots. They solve real instances of the problem by a heuristic decomposition algorithm that obtains the fleet schedule by repeatedly solving the base formulation for a single ship.

Fagerholt and Christiansen (2000a, 2000b) extend the model formulated above and study a ship scheduling problem where each ship in the fleet is equipped with a flexible cargo hold that can be partitioned into several smaller compartments in a given number of ways. The scheduling of the ships constitutes a multiship pickup and delivery problem with time windows, while the partitioning of the ships' flexible cargo holds and the allocation of cargoes to the smaller compartments is a multiallocation problem.

4.1.5 Contracted and optional cargoes

A ship scheduling problem for the tramp market boils down to pickup and delivery of cargoes at maximum profit. A tramp shipping company often engages in *Contracts of Affreightment* (COA). These are contracts to carry specified quantities of cargo between specified ports within a specific time frame for an agreed payment per ton. Mathematically, these cargoes can be handled in the same way as the cargoes for an industrial shipping problem. Tramp ships operate in a manner similar to a taxi and follow the available cargoes. They may also take optional cargoes. These optional cargoes will be picked up at a given loading port and delivered to a corresponding unloading port if the tramp shipping company finds it profitable. Thus in tramp shipping each cargo is either committed or optional and consists of a quantity given in an interval.

In the mathematical description of the problem we need to define two additional sets. For the tramp ship scheduling problem we need to partition the set of cargoes, $\mathcal{N}_{\mathrm{P}}$, into two subsets, $\mathcal{N}_{\mathrm{P}} = \mathcal{N}_{\mathrm{C}} \cup \mathcal{N}_{\mathrm{O}}$, where $\mathcal{N}_{\mathrm{C}}$ is the set of cargoes the shipping company has committed itself to carry, while $\mathcal{N}_{\mathrm{O}}$ represents the optional spot cargoes. The mathematical formulation is the same as (4.24)–(4.40), except for constraints (4.25). These constraints are split into two types of constraints as follows:

$$\sum_{v \in \mathcal{V}} \sum_{j \in \mathcal{N}_v} x_{ijv} = 1, \quad \forall i \in \mathcal{N}_{\mathrm{C}}, \tag{4.47}$$

$$\sum_{v\in\mathcal{V}}\sum_{j\in\mathcal{N}_v} x_{ijv} \leqslant 1, \quad \forall i \in \mathcal{N}_{\mathrm{O}}. \tag{4.48}$$

Constraints (4.47) ensure that all the cargoes that the shipping company has committed itself to carry are serviced. The corresponding constraints for the optional cargoes are given in (4.48). Note that the equality sign in (4.47) is replaced by an inequality in (4.48) since these cargoes do not have to be carried. When one uses a branch-and-bound algorithm to solve this problem it may be useful to insert an explicit slack variable in constraints (4.48).

A typical tramp ship scheduling problem with both optional and contracted cargoes is described in the pioneer work of Appelgren (1969, 1971). The ships in the fleet are restricted to carry only one cargo at a time, and the cargo quantities are fixed. This type of problem is extended in Brønmo et al. (2006) where cargoes are of flexible sizes for a tramp ship scheduling application.

4.1.6 Use of spot charters

In some cases the controlled fleet may have insufficient capacity to serve all cargoes for an industrial ship scheduling problem or all committed cargoes for a tramp ship scheduling problem during the planning horizon. In such a case some of the cargoes can be serviced by spot charters, which are ships chartered for a single voyage.

We extend the formulation for the tramp ship scheduling problem and introduce a variable s_i, $i \in \mathcal{N}_{\mathrm{C}}$, that is equal to 1 if cargo i is serviced by a spot charter and 0 otherwise. In addition, let π_i be the profit if cargo i is serviced by a spot charter. This profit can be either positive or negative. When we take the spot shipments into account, (4.24) and (4.25) (or (4.47)) become:

$$\max\left[\sum_{v\in\mathcal{V}}\sum_{i\in\mathcal{N}_{\mathrm{P}v}} P_i q_{iv} - \sum_{v\in\mathcal{V}}\sum_{(i,j)\in\mathcal{A}_v} C_{ijv}x_{ijv} + \sum_{i\in\mathcal{N}_{\mathrm{C}}} \pi_i s_i\right] \tag{4.49}$$

subject to

$$\sum_{v\in\mathcal{V}}\sum_{j\in\mathcal{N}_v} x_{ijv} + s_i = 1, \quad \forall i \in \mathcal{N}_{\mathrm{C}}, \tag{4.50}$$

$$s_i \in \{0, 1\}, \quad \forall i \in \mathcal{N}_{\mathrm{C}}. \tag{4.51}$$

Now, the objective function (4.49) maximizes the profit (or actually the marginal contribution, since fixed costs are excluded from the formulation). The terms are divided into the profit gained by (a) operating the fleet and (b) servicing the cargoes by spot charters. Also here it is assumed that the fleet is fixed during the planning horizon, and it is not possible to charter out some of the ships during that horizon. Constraints (4.50) ensure that all committed cargoes are serviced either by a ship in the fleet or by a spot charter. Constraints (4.51) impose the binary requirements on the spot variables. According to (4.50), these variables do not need to be defined as binary since the flow variables are binary. However, doing so might give computational advantages in a branch-and-bound process.

We can find several applications described in the literature for both tramp and industrial shipping where some of the cargoes might be serviced by spot charters, see, for instance, Bausch et al. (1998), Christiansen and Fagerholt (2002), Sherali et al. (1999), and Fagerholt (2004).

4.2 *Solution approaches for industrial and tramp scheduling models*

Theoretically the models presented in Section 4.1 can be solved directly by use of standard commercial optimization software for mixed integer linear programming after linearization of some nonlinear functions.

For instance, constraints (4.32) are given as follows:

$$x_{ijv}(l_{iv} + q_{jv} - l_{jv}) = 0, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}.$$

These constraints are linearized as

$$l_{iv} + q_{jv} - l_{jv} + V_{\mathrm{CAP}v} x_{ijv} \leqslant V_{\mathrm{CAP}v}, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}, \tag{4.52}$$

$$l_{iv} + q_{jv} - l_{jv} - V_{\mathrm{CAP}v} x_{ijv} \geqslant -V_{\mathrm{CAP}v}, \quad \forall v \in \mathcal{V}, (i,j) \in \mathcal{A}_v \mid j \in \mathcal{N}_{\mathrm{P}v}. \tag{4.53}$$

The ship capacity $V_{\mathrm{CAP}v}$ is the largest value that $(l_{iv} + q_{jv} - l_{jv})$ can take, so constraints (4.52) are redundant if x_{ijv} is equal to 0. Similarly, $(l_{iv} + q_{jv} - l_{jv})$ will never be less than $-V_{\mathrm{CAP}v}$. The schedule constraints (4.29) are linearized in the same way as constraints (4.32), but, because the original constraints have a $\leqslant$ sign, just one type of constraints is necessary in the linearized version. This way of linearizing the nonlinear constraint is also presented by Desrosiers et al. (1995).

Due to the models' complexity, only small sized data instances can be solved directly to optimality by using standard commercial optimization software. Therefore, these models usually require reformulation in order to solve them to optimality.

By studying the models presented, we see that for each cargo i we have exactly one constraint linking the ships. This corresponds to constraint types (4.2), (4.10), and (4.25) for the industrial shipping problems presented in Sections 4.1.1, 4.1.2, and 4.1.3, respectively, and constraint types (4.47) and (4.48) for the tramp shipping problems. These constraints ensure that each cargo i is served by a ship exactly once (or at most once). These constraints are called here *common constraints*. All other constraints refer to each ship v and will be called the *ship routing constraints*. For example, in the model (4.24)–(4.40), the constraints (4.26)–(4.40), constitute the routing problem for each ship where the time windows, load quantity interval and load on board the ship are considered. This observation is often exploited in the solution methods used for such type of problems. The exact solution methods are usually based on column generation approaches, where the ship routes constitute the columns. We

will therefore concentrate on two such main solution approaches, the Dantzig–Wolfe decomposition approach in Section 4.2.1, and the set partitioning approach with columns generated a priori in Section 4.2.2. Finally, in Section 4.2.3 we will briefly discuss some other approaches.

4.2.1 The Dantzig–Wolfe decomposition approach

The common constraints constitute the *master problem* in the Dantzig–Wolfe (DW) decomposition approach. None of the ship routing constraints include interaction between ships, so these constraints can be split into one *subproblem* for each ship. For each ship's subproblem, we need to find a feasible route with regard to the time windows, quantity intervals and the quantity on board the ship, so that this quantity does not exceed the capacity of the ship. Each of the feasible combinations of sailing legs (i, j) to geographical routes for a ship, including the information about starting times and load quantities at each port, is called a *ship schedule* and is indexed by r. That means a ship schedule r for ship v includes information about the values of the flow from each node i directly to node j in the geographical route, the quantity loaded or unloaded at each node i, and the starting times at each node i. The constant X_{ijvr} equals 1 if leg (i, j) by vessel v in route r and 0 otherwise. Given a geographical route, it is possible to find the optimal load quantity and starting time at each port in the route.

Since the ship routing subproblems define path structures, their extreme points correspond to paths in the underlying networks. Set $\mathcal{R}_v$ defines the extreme points for ship v. Any solution x_{ijv} satisfying the ship routing constraints can then be expressed as a nonnegative convex combination of these extreme points and must consist of binary x_{ijv} values, i.e.,

$$x_{ijv} = \sum_{r \in \mathcal{R}_v} X_{ijvr} y_{vr}, \quad \forall v \in \mathcal{V}, (i, j) \in \mathcal{A}_v, \tag{4.54}$$

$$\sum_{r \in \mathcal{R}_v} y_{vr} = 1, \quad \forall v \in \mathcal{V}, \tag{4.55}$$

$$y_{vr} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v. \tag{4.56}$$

The new variables y_{vr}, $v \in \mathcal{V}$, $r \in \mathcal{R}_v$, are called the schedule variables, and equal 1 if ship v chooses to sail schedule r. Let $A_{ivr} = \sum_{j \in \mathcal{N}_v} X_{ijvr}$ be equal to 1 if schedule r for ship v services cargo i and 0 otherwise. The column vector in the master problem contains information about the actual cargoes in schedule r for ship v. In addition, the optimal geographical route, the arrival times and the size of the cargoes for the given set of cargoes for a schedule (v, r) determine the profit coefficient in the objective function for the corresponding column.

The master problem in the DW decomposition approach. Substituting (4.54)–(4.56) in (4.24) and (4.25), the integer master problem for the industrial ship

scheduling problem with flexible cargo sizes is transformed into:

$$\max \sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} P_{vr} y_{vr} \tag{4.57}$$

subject to

$$\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} A_{ivr} y_{vr} = 1, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, \tag{4.58}$$

$$\sum_{r \in \mathcal{R}_v} y_{vr} = 1, \quad \forall v \in \mathcal{V}, \tag{4.59}$$

$$y_{vr} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v. \tag{4.60}$$

The objective function (4.57) maximizes the profit, where P_{vr} is the profit of carrying the cargoes on schedule r by ship v, respectively. Constraints (4.58) ensure that all cargoes are serviced by a ship in the company's fleet. Constraints (4.59) assure that each ship in the fleet is assigned exactly one schedule. Constraints (4.60) impose the binary requirements on the variables.

The corresponding master problem for the tramp ship routing and scheduling problem with spot charters can be formulated as follows:

$$\max \left[\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} P_{vr} y_{vr} + \sum_{i \in \mathcal{N}_{\mathrm{C}}} \pi_i s_i \right] \tag{4.61}$$

subject to

$$\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} A_{ivr} y_{vr} + s_i = 1, \quad \forall i \in \mathcal{N}_{\mathrm{C}}, \tag{4.62}$$

$$\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} A_{ivr} y_{vr} \leqslant 1, \quad \forall i \in \mathcal{N}_{\mathrm{O}}, \tag{4.63}$$

$$\sum_{r \in \mathcal{R}_v} y_{vr} = 1, \quad \forall v \in \mathcal{V}, \tag{4.64}$$

$$y_{vr} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v, \tag{4.65}$$

$$s_i \in \{0, 1\}, \quad \forall i \in \mathcal{N}_{\mathrm{C}}. \tag{4.66}$$

Column generation and the subproblems within the DW decomposition approach. The models (4.57)–(4.60) and (4.61)–(4.66) are based on knowledge of all feasible ship schedules (columns). However, for some real ship scheduling problems it is time consuming to generate all these schedules, and the number of such schedules would result in too many columns when solving the models. Instead, we solve the LP-relaxation of the *restricted* master problem which only differs from the continuous original master problem by having fewer variables. First, an initial restricted master problem is solved. Then some new columns are added to the restricted master problem. These columns correspond to ship schedules with positive reduced costs in the solution of the (maximization)

master problem. This means that the dual values from the solution of the restricted master problem are transferred to the subproblems. The subproblems are solved and ship schedules are generated. The restricted master problem is reoptimized with the added new columns, resulting in new dual values. This procedure continues until no columns with positive reduced costs exist, and no improvements can be made. At that point all the feasible solutions in the original master problem have been implicitly evaluated. A continuous optimal solution is then attained for both the original and the restricted master problem. This LP-relaxed solution approach can be embedded in a branch-and-bound search to find an optimal solution.

The subproblems can be formulated as shortest path problems and solved by specific dynamic programming algorithms on generated networks for each ship. The underlying network for each ship is specified by nodes, each of which includes information about the port and the corresponding cargo with time window for starting service and feasible cargo quantities. The recursive formulas in the dynamic programming algorithms include the expressions for the reduced costs. Algorithms for solving such problems are thoroughly described in Desrosiers et al. (1995) and, for a special ship scheduling problem, in Christiansen and Nygreen (1998b).

The DW decomposition approach has been used in numerous vehicle routing applications during the last twenty years. However, Appelgren (1969, 1971) was the first one to use this approach for a pickup and delivery problem with time windows, and that application was for the tramp shipping industry. Another ship routing application using the DW decomposition approach was studied by Christiansen (1999) (see also Christiansen and Nygreen, 1998a, 1998b) and is discussed in Section 4.3.1.

4.2.2 The set partitioning approach

Ship scheduling problems are often tightly constrained, and in such a case it is possible to generate schedules for all cargo combinations for all ships (i.e., all columns) a priori. The original arc flow models given in Section 4.1 can be transformed to path flow models, and these path flow models correspond to the master problems (4.57)–(4.60) and (4.61)–(4.66) in the Dantzig–Wolfe (DW) decomposition approach. Both models are set partitioning (SP) models or can easily be transformed into a SP model by introducing a slack variable to constraints (4.63). In this approach all column vectors for the set partitioning model are generated in advance, and a binary variable y_{vr} is defined for each column vector generated. We can find numerous ship scheduling applications where this approach is used, see for instance Brown et al. (1987), Fisher and Rosenwein (1989), Bausch et al. (1998), Fagerholt (2001), Fagerholt and Christiansen (2000a, 2000b), Christiansen and Fagerholt (2002), and Brønmo et al. (2006).

Here, we are generating columns for all feasible cargo combinations for a particular ship v. For each of the feasible cargo combinations, we have to find the geographical route, arrival time at the ports and the load quantities of the

cargoes, such that the sum of the profits in the schedule is maximized. Further, each node has to be serviced within its specified load interval and time window. Finally, the loading node has to be visited before its corresponding unloading node. If the ships in the fleet are equipped with cargo holds or tanks of various capacities, the optimal allocation of products to tanks has to be determined as well. All constraints that are exclusive for a particular ship have to be considered in the column generation phase of this approach. The problem of finding the optimal route and schedule for a single ship can be solved by using dynamic programming or by enumerating all feasible combinations of routes for a given set of cargoes. Both approaches have been used. Fagerholt and Christiansen (2000b) describe a dynamic programming approach for a combined multiship pickup and delivery problem with time windows and a multiallocation problem, while Brønmo et al. (2006) describes an enumeration procedure for a tramp scheduling problem with flexible cargo sizes.

4.2.3 Other solution approaches

In general, many solution methods, both optimization-based and heuristic ones, were developed to solve routing and scheduling problems for other modes of transportation. These methods can often be used with some minor modifications for ship scheduling problems. Here we report several studies in the ship scheduling literature where solution approaches other than the ones discussed in Sections 4.2.1 and 4.2.2 were used.

Sherali et al. (1999) presented an aggregated mixed integer programming model retaining the principal features of the real ship scheduling problem with various cargo hold capacities and possible spot charters. A rolling horizon heuristic is developed to solve the problem.

The ship scheduling problem studied by Scott (1995) is solved by applying Lagrangian relaxation to the model to produce a set of potentially good schedules, containing the optimal cargo schedule. A novel refinement of Benders' decomposition is then used to choose the optimum schedule from within the set, by avoiding solving an integer LP-problem at each iteration. The method manages to break a difficult integer programming (IP) problem into two relatively simple steps which parallel the steps typically taken by schedulers.

The tramp ship scheduling problem is studied by Brønmo et al. (2006, 2007), and two solution approaches are suggested and compared. In addition to a set partitioning approach, they describe a multistart heuristic consisting of two phases. First multiple initial solutions are constructed by a simple insertion method. Then a subset of the best initial solutions is improved by a quick local search. A few of the best resulting solutions from the quick local search are improved by an extended local search.

4.3 Maritime supply chains

A maritime supply chain is a supply chain where sea transport constitutes at least one vital link. Supply chains of companies with foreign sources of raw

materials or with overseas customers very often include maritime transportation. Supply chain optimization is an active field of research, and we can see applications in almost all industries. However, the focus of such applications is usually not on maritime transportation. At the tactical planning level the supply chain perspective is missing in ship routing and scheduling studies reported in the literature.

Fleet scheduling is often performed under tight constraints. The shipper specifies the cargoes with little or no flexibility in cargo quatnities and the time widows are unnecessarily tight. The shipping company tries to find an optimal fleet schedule based on such requirements while trying to maximize the profit (or minimizing the costs). Realizing the potential of relaxing such constraints, Brønmo et al. (2006) and Fagerholt (2001) considered flexibility in shipment sizes and in time windows. The results of their studies show that there might be a great potential in collaboration and integration along the supply chain, for instance between the shippers and the shipping company.

Vendor managed inventory (VMI) takes advantage of the benefits of introducing flexibility in delivery time windows and cargo quantities, and transfers inventory management and ordering responsibilities completely to the vendor or the logistics provider. From recent literature and from our active contacts with the shipping industry we see that an increased number of shipping companies play the role of vendors in such logistics systems.

In this section we emphasize combined ship scheduling and inventory management problems in the industrial and tramp shipping sectors. Section 4.3.1 discusses such a problem for transportation of a single product, while Section 4.3.2 considers planning problems with multiple products. Finally, in Section 4.3.3 we will comment on some other research within supply chain optimization that focuses on ship scheduling.

4.3.1 Inventory routing for a single product

In industrial maritime transportation, the transporter has often a twofold responsibility. In this segment large quantities are transported, and normally considerable inventories exist at each end of a sailing leg. In some situations, the transporter has both the responsibility for the transportation and the inventories at the sources and at the destinations. We consider a planning problem where a single product is transported, and we call this problem the single product inventory ship routing problem (s-ISRP). The single product is produced at the sources, and we call the associated ports loading ports. Similarly, the product is consumed at certain destinations and the corresponding ports are called unloading ports. Inventory storage capacities are given in all ports, and the planners have information about the production and consumption rates of the transported product. We assume that these rates are constant during the planning horizon. In contrast to most ship scheduling problems, the number of calls at a given port during the planning horizon is not predetermined, neither is the quantity to be loaded or unloaded in each port call. The production or consumption rate and inventory information at each port, together with ship

capacities and the location of the ports, determine the number of possible calls at each port, the time windows for start of service and the range of feasible load quantities for each port call.

If the product is loaded and unloaded in time at the sources and destinations, respectively, neither production nor consumption will be interrupted. The planning problem is therefore to find routes and schedules that minimize the transportation cost without interrupting production or consumption. The transporter owns both the producing sources and consuming destinations and controls the inventories at both ends, so the inventory costs do not come into play. The transporter operates a heterogeneous fleet of ships.

This s-ISRP has many similarities to the ship scheduling problem with flexible cargo sizes. In contrast to the problem described in Section 4.1.3, the number of cargoes is not given in advance, neither is the number of ship calls at a port. Further, we have no predetermined loading and unloading port for a particular cargo. In contrast to the problem described in Section 4.1.3, we assume that the ship is not necessarily empty in the beginning of the planning horizon but might have some load onboard. In addition, we have to keep track of the inventory levels. There must be sufficient product in consumption inventories, and their inventory in production ports cannot exceed the inventory storage capacity. In addition, storage capacity limits exist for all consumption inventories.

In the mathematical description of the problem each port is represented by an index i and the set of ports is given by $\mathcal{N}$. Let $\mathcal{V}$, indexed by v, be the set of available ships to be routed and scheduled. Not all ships can visit all ports, and $\mathcal{N}_v = \{\text{feasible ports for ship } v\} \cup \{o(v), d(v)\}$ is the set of ports that can be visited by ship v. The terms $o(v)$ and $d(v)$ represent the artificial origin port and artificial destination port of ship v, respectively. Each port can be visited several times during the planning horizon, and $\mathcal{M}_i$ is the set of possible calls at port i, while $\mathcal{M}_{iv}$ is the set of calls at i that can be made by ship v. The port call number is represented by an index m, and M_i is the last possible call at port i. The necessary calls to a port are given by the set $\mathcal{M}_{Ci}$ and these necessary calls have similarities to the contracted cargoes in the problems discussed in Section 4.1.5.

The set of nodes in the flow network represents the set of port calls, and each port call is specified by (i, m), $i \in \mathcal{N}$, $m \in \mathcal{M}_i$. In addition, we specify flow networks for each ship v with nodes (i, m), $i \in \mathcal{N}_v$, $m \in \mathcal{M}_{iv}$. Finally, $\mathcal{A}_v$ contains all feasible arcs for ship v, which is a subset of $\{i \in \mathcal{N}_v, m \in \mathcal{M}_{iv}\} \times \{i \in \mathcal{N}_v, m \in \mathcal{M}_{iv}\}$.

Figure 8 shows an artificial, simplified case consisting of five ports and two ships. Each potential port call is indicated by a node. We see that port 1 can be called three times during the planning horizon. We have three loading ports and two unloading ports. The arrows indicate a solution to the planning problem where the routes and schedules satisfy the time windows and inventory constraints.

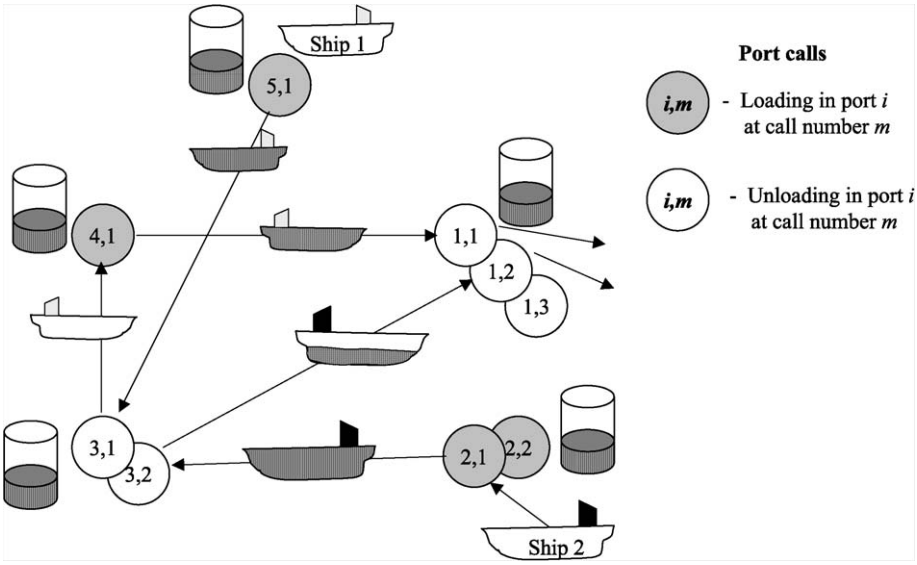


Fig. 8. A solution for a single product inventory routing problem with 5 ports and 2 ships.

Port 5 is the initial location for ship 1. The ship loads up to its capacity before sailing to port call (3, 1) and unloading this quantity. The ship continues to port call (4, 1) to load before ending up at port call (1, 1). Ship 2 is empty at sea at the beginning of the planning horizon and starts service at port call (2, 1) after some time. Here the ship loads to its capacity before sailing toward port call (3, 2). At port call (3, 2) the ship unloads half of its load before it continues to port call (1, 2) and unloads the rest of the quantity on board. Here, two unloading ports are called in succession.

Port 3 is called several times during the planning horizon. The solid, gray line in Figure 9 shows the inventory level for port 3 during the planning horizon. Ship 2 unloads half of its load at port call (3, 2) as soon as possible. Here it is important to ensure that the inventory level does not exceed the maximal one when the unloading ends. Regardless of the rest of the planning problem, the broken line in Figure 9 illustrates another extreme situation where ship 2 starts the service at port 3 as late as possible. Here, the inventory level is not allowed to be under the minimal stock level when the unloading starts. From these two extreme scenarios for the inventory levels, we can derive the feasible time window for port call (3, 2) given that the rest of the planning problem remains unchanged.

The variable quantity interval is given by $[Q_{\mathrm{MN}im}, Q_{\mathrm{MX}imv}]$, where $Q_{\mathrm{MN}im}$ is the minimum quantity to be (un)loaded at port call (i, m) given that the port is called, while $Q_{\mathrm{MX}imv}$ is the maximum quantity to be (un)loaded at port call (i, m) for ship v. The capacity of ship v is given by $V_{\mathrm{CAP}v}$.

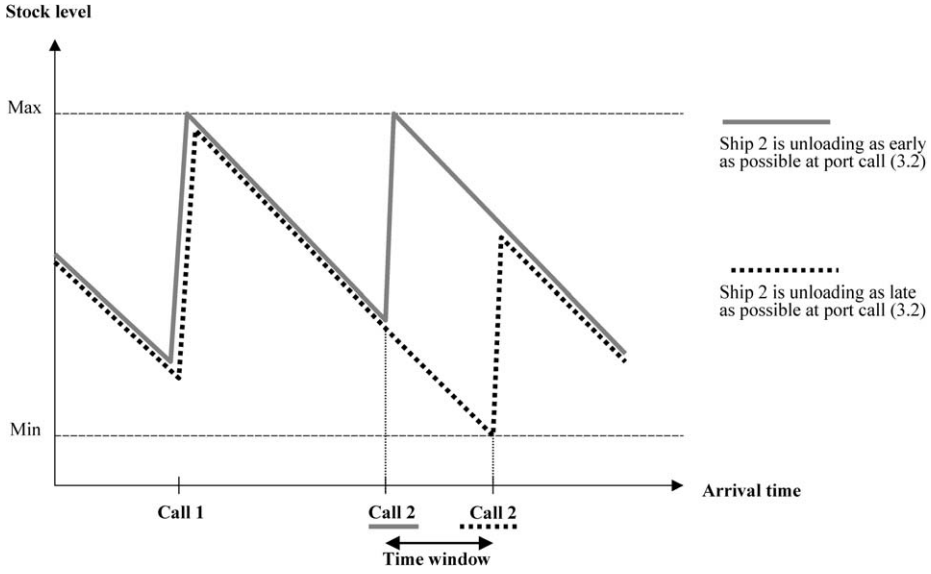


Fig. 9. The inventory level at port 3 during the planning horizon.

The time required to load or unload one unit of a cargo at port i is given by T_{Qi}. The term T_{Sijv} represents the sailing time from port i to port j with ship v. Let $[T_{MNim}, T_{MXim}]$ denote the arrival time window associated with port call (i, m). This time window can be calculated based on other data in the model, such as the inventory conditions. In addition, for some port calls the time windows are explicitly given. In a preprocessing phase, it is important to make efforts to reduce the time window widths. In some ports, there is a minimum required time, T_{Bi}, between a departure of one ship and the arrival of the next ship, due to small port area or narrow channels from the port to the pilot station. Let T denote the planning horizon.

The levels of the inventory (or stock) have to be within a given interval at each port $[S_{MNi}, S_{MXi}]$. The production rate R_i is positive if port i is producing the product, and negative if port i is consuming the product. Further, constant I_i is equal to 1, if i is a loading port, -1, if i is an unloading port, and 0, if i is $o(v)$ or $d(v)$. The total variable cost C_{ijv} that includes port, channel, and fuel oil costs, corresponds to a sailing from port i to port j with ship v.

In the mathematical formulation we use the following types of variables: the binary flow variable x_{imjnv}, $v \in \mathcal{V}$, $(i, m, j, n) \in \mathcal{A}_v$, equals 1, if ship v sails from node (i, m) directly to node (j, n), and 0 otherwise, and the slack variables w_{im}, $i \in \mathcal{N}$, $m \in \mathcal{M}_i \backslash \mathcal{M}_{Ci}$, is equal to 1 if no ship takes port call (i, m), and 0 otherwise. The time variable t_{im}, $(i \in \mathcal{N}, m \in \mathcal{M}_i) \cup (i \in o(v), \forall v, m = 1)$, represents the time at which service begins at node (i, m). Variable l_{imv}, $v \in V$, $i \in \mathcal{N}_v \backslash \{d(v)\}$, $m \in \mathcal{M}_{iv}$, gives the total load onboard ship v just after the service is completed at node (i, m), while variable q_{imv}, $v \in V$, $i \in \mathcal{N}_v \backslash \{d(v)\}$,

$m \in \mathcal{M}_{iv}$, represents the quantity loaded or unloaded at port call (i, m), when ship v visits (i, m). Finally, s_{im}, $i \in \mathcal{N}$, $m \in \mathcal{M}_i$, represents the inventory (or stock) level when service starts at port call (i, m). It is assumed that nothing is loaded or unloaded at the artificial origin $o(v)$ and that the ship arrives at $o(v)$ at a given fixed time; $t_{o(v)1} = T_{\mathrm{MN}o(v)1} = T_{\mathrm{MX}o(v)1}$. The ships may have cargo onboard, L_{0v}, at the beginning of the planning horizon; $l_{o(v)1v} = L_{0v}$. At the beginning of the planning horizon, the stock level at each port i is S_{0i}.

The arc flow formulation of the single product inventory ship routing problem (s-ISRP) is as follows:

$$\min \sum_{v \in \mathcal{V}} \sum_{(i,m,j,n) \in \mathcal{A}_v} C_{ijv} x_{imjnv} \tag{4.67}$$

subject to

$$\sum_{v \in \mathcal{V}} \sum_{j \in \mathcal{N}_v} \sum_{n \in \mathcal{M}_{jv}} x_{imjnv} + w_{im} = 1, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.68}$$

$$\sum_{j \in \mathcal{N}_v} \sum_{n \in \mathcal{M}_{jv}} x_{o(v)1jnv} = 1, \quad \forall v \in \mathcal{V}, \tag{4.69}$$

$$\sum_{i \in \mathcal{N}_v} \sum_{m \in \mathcal{M}_{iv}} x_{imjnv} - \sum_{i \in \mathcal{N}_v} \sum_{m \in \mathcal{M}_{iv}} x_{jnimv} = 0,$$
$$\forall v \in \mathcal{V}, j \in \mathcal{N}_v \backslash \{o(v), d(v)\}, n \in \mathcal{M}_{jv}, \tag{4.70}$$

$$\sum_{i \in \mathcal{N}_v} \sum_{m \in \mathcal{M}_{iv}} x_{imd(v)1v} = 1, \quad \forall v \in \mathcal{V}, \tag{4.71}$$

$$x_{imjnv}(t_{im} + T_{\mathrm{Q}i} q_{imv} + T_{\mathrm{S}ijv} - t_{jn}) \leqslant 0,$$
$$\forall v \in \mathcal{V}, (i, m, j, n) \in \mathcal{A}_v \mid j \neq d(v), \tag{4.72}$$

$$t_{o(v)1} = T_{\mathrm{MN}o(v)1} = T_{\mathrm{MX}o(v)1}, \quad \forall v \in \mathcal{V}, \tag{4.73}$$

$$T_{\mathrm{MN}im} \leqslant t_{im} \leqslant T_{\mathrm{MX}im}, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.74}$$

$$x_{imjnv}(l_{imv} + I_j q_{jnv} - l_{jnv}) = 0,$$
$$\forall v \in \mathcal{V}, (i, m, j, n) \in \mathcal{A}_v \mid j \neq d(v), \tag{4.75}$$

$$q_{o(v)1v} = 0, \quad \forall v \in \mathcal{V}, \tag{4.76}$$

$$l_{o(v)1v} = L_{0v}, \quad \forall v \in \mathcal{V}, \tag{4.77}$$

$$q_{imv} \leqslant l_{imv} \leqslant \sum_{j \in \mathcal{N}_v} \sum_{n \in \mathcal{M}_{jv}} V_{\mathrm{CAP}v} x_{imjnv},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}_v, m \in \mathcal{M}_{iv} \mid I_i = 1, \tag{4.78}$$

$$0 \leqslant l_{imv} \leqslant \sum_{j \in \mathcal{N}_v} \sum_{n \in \mathcal{M}_{jv}} V_{\mathrm{CAP}v} x_{imjnv} - q_{imv},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}_v, m \in \mathcal{M}_{iv} \mid I_i = -1, \tag{4.79}$$

$$q_{imv} \leqslant \sum_{j \in \mathcal{N}_v} \sum_{n \in \mathcal{M}_{jv}} Q_{\mathrm{MX}imv} x_{imjnv},$$
$$\forall v \in \mathcal{V}, i \in \mathcal{N}_v \backslash \{o(v), d(v)\}, m \in \mathcal{M}_{iv}, \tag{4.80}$$
$$\sum_{v \in \mathcal{V}} q_{imv} + Q_{\mathrm{MN}im} w_{im} \geqslant Q_{\mathrm{MN}im}, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.81}$$
$$s_{i1} - R_i t_{i1} = S_{0i}, \quad \forall i \in \mathcal{N}, \tag{4.82}$$
$$s_{i(m-1)} - \sum_{v \in \mathcal{V}} I_i q_{i(m-1)v} + R_i(t_{im} - t_{i(m-1)}) - s_{im} = 0,$$
$$\forall i \in \mathcal{N}, m \in \mathcal{M}_i \backslash \{1\}, \tag{4.83}$$
$$S_{\mathrm{MN}i} \leqslant s_{im} \leqslant S_{\mathrm{MX}i}, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.84}$$
$$S_{\mathrm{MN}i} \leqslant s_{im} - \sum_{v \in \mathcal{V}} I_i q_{imv} + R_i(T - t_{im}) \leqslant S_{\mathrm{MX}i},$$
$$\forall i \in \mathcal{N}, m = M_i, \tag{4.85}$$
$$w_{im} - w_{i(m-1)} \geqslant 0, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i \backslash \mathcal{M}_{\mathrm{C}i}, \tag{4.86}$$
$$t_{im} - t_{i(m-1)} - \sum_{v \in \mathcal{V}} T_{\mathrm{Q}i} q_{i(m-1)v} + T_{\mathrm{B}i} w_{im} \geqslant T_{\mathrm{B}i},$$
$$\forall i \in \mathcal{N}, m \in \mathcal{M}_i \backslash \{1\}, \tag{4.87}$$
$$x_{imjnv} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, (i, m, j, n) \in \mathcal{A}_v, \tag{4.88}$$
$$w_{im} \in \{0, 1\}, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i \backslash \mathcal{M}_{\mathrm{C}i}. \tag{4.89}$$

The objective function (4.67) minimizes the total costs. Constraints (4.68) ensure that each port call is visited at most once. Constraints (4.69)–(4.71) describe the flow on the sailing route used by ship v. Constraints (4.72) take into account the timing on the route. Initial time conditions for each ship are defined by constraints (4.73). The time windows are given by constraints (4.74). If no ship is visiting port call (i, m), we will get an artificial start time within the time windows for a "dummy ship". These artificial start times are used in the inventory balances. Constraints (4.75) give the relationship between the binary flow variables and the ship load at each port call. Initial conditions for the load quantity and the quantity on board are given in constraints (4.76) and (4.77), respectively. Constraints (4.78) and (4.79) give the ship capacity intervals at the port calls for loading and unloading ports, respectively. Constraints (4.80) and (4.81) are the load limit constraints. All constraints (4.68)–(4.81) so far are similar to constraints (4.25)–(4.37) for the industrial ship scheduling problem with flexible cargo sizes in Section 4.1.3. In addition, we have some inventory constraints for this problem. The inventory level at the first call in each port is calculated in constraints (4.82). From constraints (4.83), we find the inventory level at any port call (i, m) from the inventory level upon arrival at the port in the previous call $(i, m - 1)$, adjusted for the loaded/unloaded quantity at the port call and the production/consumption between the two arrivals. The general inventory limit constraints at each port call are given in (4.84).

Constraints (4.85) ensure that the level of inventory at the end of the planning horizon is within its limits. It can be easily shown by substitution that constraints (4.85) ensure that the inventory at time T will be within the bounds even if ports are not visited at the last calls. One or several of the calls in a specified port can be made by a dummy ship, and the highest call numbers will be assigned to dummy ships in constraints (4.86). These constraints reduce the number of symmetrical solutions in the solution approach. For the calls made by a dummy ship, we get artificial starting times within the time windows and artificial stock levels within the inventory limits. Constraints (4.87) prevent service overlap in the ports and ensure the order of real calls in the same port. A ship must complete its service before the next ship starts its service in the same port. Finally, the formulation involves binary requirements (4.88) and (4.89) on the flow variables and port call slack variables, respectively.

This s-ISRP can be solved by the Dantzig–Wolfe (DW) decomposition approach described in Section 4.2.1, where we have a ship routing and scheduling problem for each ship and an inventory management problem for each port. However, if we try to decompose the model directly, it does not separate due to the starting time t_{im} and the load quantity q_{imv} variables. These variables are needed in both subproblems that we have here, the routing and the inventory subproblems. This issue is resolved by introducing new time and quantity variables, such that we get variables for each (i, m, v)-combination (t_{imv} and q_{imv}) and each port call (t_{im} and q_{im}) and introducing coupling constraints to the problem as follows:

$$(1 - w_{im})\left[t_{im} - \sum_{v \in \mathcal{V}} t_{imv}\right] = 0, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.90}$$

$$q_{im} - \sum_{v \in \mathcal{V}} q_{imv} = 0, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i. \tag{4.91}$$

Now, the constraint set can be split into three independent groups. The first constraint group consists of ship constraints and constitutes the routing problem for each ship where the time windows and load on board the ship are considered. The *ship routing constraints* are based on constraints (4.69)–(4.81) with the starting time, t_{imv}, and load quantity, q_{imv}, variables. The *port inventory constraints* describe the inventory management problem for each port, and here t_{im} and q_{im} are used in the problem and are based on constraints (4.74) and (4.80)–(4.87). The remaining constraints are the *common constraints* (4.68), (4.90), and (4.91).

As described in Section 4.2.1 we introduce a variable y_{vr} for each of the feasible combinations of sailing legs to geographical routes, starting times and load quantities at the port calls, and such a combination is called a ship schedule r, $r \in \mathcal{R}_v$. The schedule r includes information about the sailed legs in the route (X_{imjnvr} equals 0 or 1), number of visits at port call (A_{imvr} equals 0 or 1), the load quantity of each port call ($Q_{\mathrm{V}imvr}$), and the starting time of each port

call ($T_{\mathrm{V}imvr}$). No quantity and starting time information is given for "dummy calls".

At the ports, it is important to determine the load quantity and starting time at each call in the port such that the inventory level is within its limits during the entire planning horizon. Each of the feasible combinations of load quantities, starting times and number of calls at a port i during the planning horizon is called a *port call sequence* s, $s \in \mathcal{S}_i$. The values of $Q_{\mathrm{H}ims}$ and $T_{\mathrm{H}ims}$ represent the load quantity and starting time for the port call (i, m) in sequence s, respectively. The value of W_{ims} is 1 if sequence s is not visiting port call (i, m), and from this constant we can find the number of calls at port i. Let variable z_{is}, $i \in \mathcal{N}$, $s \in \mathcal{S}_i$, be 1, if port i selects sequence s and 0 otherwise.

The resulting master problem becomes:

$$\min \sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} C_{vr} y_{vr} \tag{4.92}$$

subject to

$$\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} A_{imvr} y_{vr} + \sum_{s \in S_i} W_{ims} z_{is} = 1, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.93}$$

$$\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} Q_{\mathrm{V}imvr} y_{vr} - \sum_{s \in S_i} Q_{\mathrm{H}ims} z_{is} = 0, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.94}$$

$$\sum_{v \in \mathcal{V}} \sum_{r \in \mathcal{R}_v} T_{\mathrm{V}imvr} y_{vr} - \sum_{s \in S_i} T_{\mathrm{H}ims} z_{is} = 0, \quad \forall i \in \mathcal{N}, m \in \mathcal{M}_i, \tag{4.95}$$

$$\sum_{r \in \mathcal{R}_v} y_{vr} = 1, \quad \forall v \in \mathcal{V}, \tag{4.96}$$

$$\sum_{s \in \mathcal{S}_i} z_{is} = 1, \quad \forall i \in \mathcal{N}, \tag{4.97}$$

$$y_{vr} \geqslant 0, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v, \tag{4.98}$$

$$z_{is} \geqslant 0, \quad \forall i \in \mathcal{N}, s \in \mathcal{S}_i, \tag{4.99}$$

$$\sum_{r \in \mathcal{R}_v} X_{imjnvr} y_{vr} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, (i, m, j, n) \in \mathcal{A}_v. \tag{4.100}$$

The objective function (4.92) minimizes the transportation costs. No such costs exist for the inventory problem, so just the route variables with associated costs are present. Unlike usual vehicle routing problems solved by a DW decomposition approach, the master problem includes additional coupling constraints for the load quantities and starting times to synchronize the port inventory and ship route aspects. These are given in constraints (4.94) and

(4.95), respectively. The convexity rows for the ships and ports are given in constraints (4.96) and (4.97). The integer requirements are defined by (4.100) and correspond to declaring the original flow variables x_{imjnv} as binary variables.

In the DW decomposition approach, the port call sequences and ship schedules with least reduced costs in the (minimization) master problem are generated. This procedure is described in Section 4.2.1 for a maximization problem. We solve subproblems for each port and each ship, and both types of subproblems can be solved by dynamic programming algorithms. Christiansen (1999) studies a real ship scheduling and inventory management problem for transportation of ammonia. The overall solution approach is described in Christiansen and Nygreen (1998a), and the method for solving the subproblems is given in detail in Christiansen and Nygreen (1998b).

In the real problem described by Christiansen (1999), the shipper trades ammonia with other operators in order to better utilize the fleet and to ensure the ammonia balance at it's own plants. These traded volumes are determined by negotiations. The transporter undertakes to load or unload ammonia within a determined quantity interval and to arrive at a particular external port within a given time window. For these external ports, no inventory management problem exists. This is an example of a shipper operating its fleet in both the industrial and tramp modes simultaneously.

Another solution approach to the same problem was developed by Flatberg et al. (2000). They have used an iterative improvement heuristic combined with an LP solver to solve this problem. The solution method presented consists of two parts. Their heuristic is used to solve the combinatorial problem of finding the ship routes, and an LP model is used to find the starting time of service at each call and the loading or unloading quantity. Computational results for real instances of the planning problem are reported. However, no comparisons in running time or solution quality of the results in Flatberg et al. (2000) and Christiansen and Nygreen (1998a) exist.

At the unloading ports ammonia is further processed into different fertilizer products, and these products are supplied to the agricultural market. Fox and Herden (1999) describe a MIP model to schedule ships from such ammonia processing plants to eight ports in Australia. The objective is to minimize freight, discharge and inventory holding costs while taking into account the inventory, minimum discharge tonnage and ship capacity constraints. Their multiperiod model is solved by a commercial optimization software package.

4.3.2 Inventory routing for multiple products

When there are multiple products, the inventory ship routing problem becomes much harder to solve. Until recently this problem was scarcely considered in the literature. However, Hwang (2005) studied this problem in his PhD thesis and assumed that the various products require dedicated compartments in the ship. The problem studied is to decide how much of each product should be carried by each ship from production ports to consuming ports, subject to the inventory level of each product in each port being maintained between cer-

tain levels. These levels are set by the production/consumption rates and the storage capacities of the various products in each port. The problem is formulated as a mixed-integer linear programming problem with a special structure. Small test problems are randomly generated and solved. Hwang uses a combined Lagrangian relaxation and heuristic approach to solve the test problems.

In this section, we consider a special case of the multiple inventory routing problem where several products are produced at several plants located adjacent to ports, and the same products are consumed at consuming plants in other ports. In contrast to the single inventory ship routing problem (s-ISRP) described in Section 4.3.1, we assume that in the problem considered here, the shipper does not control and operate the fleet of ships. The transportation is carried out by ships that are chartered for performing single voyages from a loading to an unloading port at known cost (spot charters). This means that the focus of the problem is to optimally determine the quantity and timing of shipments to be shipped, while the routing of the ships is not an important part of the problem.

As before, we call the production plants loading ports and the consuming plants unloading ports. Not all the products are produced or consumed at all the plants. The plants have limited storage capacity for the products that they produce or consume. Unlike the s-ISRP discussed in Section 4.3.1, the production and consumption rates may vary over time. However, total production and total consumption of each product are balanced over time. It is therefore possible to produce and consume continuously at all the plants while the inventories are between their lower and upper limits, given that the products are shipped from the loading ports to the unloading ports frequently enough. Prevention of interruption in production or consumption at all plants due to lack of materials or storage space is the main goal of our planning, same as for the s-ISRP.

Ship voyages have a single loading port and a single unloading port. We assume that the cost of a voyage between two ports consists of two components, a fixed set-up cost, and a variable cost per unit shipped that is based on the distance between the two ports. Further, we assume that there is a sufficient number of ships of different sizes. Figure 10 illustrates the situation modeled, where the bold arcs are in the model and the stippled ones are not.

There is uncertainty both in the sailing times and in the production and consumption rates. This is taken into account by the use of safety stocks in the inventory planning. If a ship arrives late at a loading port, production may stop at the plant due to shortage of storage space for the products. To reduce the possibility of such situations, the storage capacities modeled are less than the actual capacities. This has the same effect as the use of safety stocks. In our model we set an upper safety stock level that is below the storage capacity, and a lower safety stock level that is above a specified lower storage capacity. Any diversion of the inventory from this band of safety stock limits is penalized, as illustrated in Figure 11.

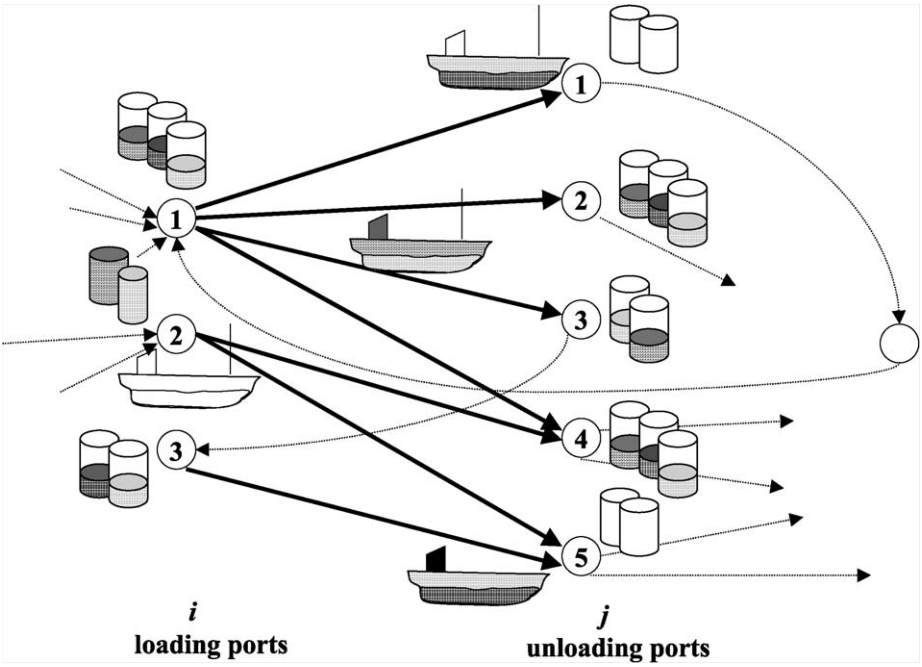


Fig. 10. A multiple product inventory routing problem. The bold arcs are in the model, the stippled ones are not.

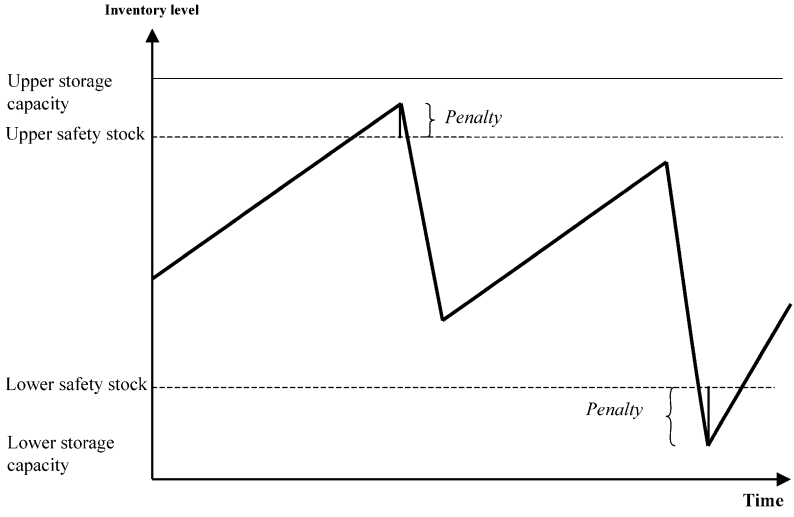


Fig. 11. The inventory level during the planning horizon for one port.

In Figure 11, we see the inventory level during the planning horizon for one of the products produced at a loading port. The port is visited twice during the planning horizon. The production rate is lower than the ship loading rate. Compared with the s-ISRP where time was continuous, we revert here to using one day (24 hours) time increments. We measure the transportation time in days such that all products produced in one day can leave the loading port on the same day, and all products that arrive at an unloading port during a day can be consumed on the same day. However, introducing a one day lag between these operations requires only a minor change in the formulation. Further, we assume at most one ship sailing per day between any loading and unloading port pair.

The objective of the model is to find a transportation plan that minimizes the sum of the transportation cost and the inventory penalties.

In the mathematical description of the problem, let $\mathcal{N}$ be the set of ports indexed by i or j. Divide this set into the subset of loading ports $\mathcal{N}_{\mathrm{P}}$ and the subset of unloading ports $\mathcal{N}_{\mathrm{D}}$. Let $\mathcal{K}$ be the set of products indexed by k, and let $\mathcal{T}$ be the set of periods (days) indexed by t.

The time for sailing from loading port i to unloading port j including the loading and unloading time is T_{ij}. R_{ikt} is production or consumption of product k in port i during day t. These rates are positive in loading ports and negative in unloading ports.

The inventory information is given by the storage capacities and the safety stock. The absolute lower and upper storage capacities for product k in port i are 0 and $S_{\mathrm{MX}ik}$, respectively. The lower and upper safety stocks for the same products in the same ports are $S_{\mathrm{SL}ik}$ and $S_{\mathrm{SU}ik}$. The inventory in the beginning of the planning horizon for product k in port i is given by $S_{\mathrm{ST}ik}$.

U_{ij} represents the maximal capacity/size of a ship that can sail between the loading port i and unloading port j. Due to the setup cost involved with a voyage between two ports the transportation cost will be minimized by using the largest ship possible. By always using ships of maximal size, the model becomes simple.

The fixed cost for sailing a ship from loading port i to unloading port j is represented by $C_{\mathrm{F}ij}$, while $C_{\mathrm{V}ij}$ is the variable cost of shipping one ton of a product between i and j. The penalty cost per day for each ton of lower (upper) safety stock shortfall (excess) for product k in port i is $C_{\mathrm{L}ik}$ ($C_{\mathrm{U}ik}$).

In the mathematical formulation we use the following types of variables: the binary flow variable x_{ijt}, $i \in \mathcal{N}_{\mathrm{P}}$, $j \in \mathcal{N}_{\mathrm{D}}$, $t \in \mathcal{T}$, equals 1, if a ship sails from port i on day t to port j, and 0 otherwise. The quantity variable q_{ijkt}, $i \in \mathcal{N}_{\mathrm{P}}$, $j \in \mathcal{N}_{\mathrm{D}}$, $k \in \mathcal{K}$, $t \in \mathcal{T}$, represents the number of tons of product k that leaves port i on day t on a ship bounded for port j. The inventory of product k at the end of day t in port i is given by s_{ikt}, $i \in \mathcal{N}$, $k \in \mathcal{K}$, $t \in \mathcal{T}$, while the lower safety stock shortfall and the upper safety stock excess of product k at the end of day t in port i are $s_{\mathrm{L}ikt}$, $i \in \mathcal{N}$, $k \in \mathcal{K}$, $t \in \mathcal{T}$, and $s_{\mathrm{U}ikt}$, $i \in \mathcal{N}$, $k \in \mathcal{K}$, $t \in \mathcal{T}$, respectively.

The mathematical formulation of the multiple product inventory ship routing problem considered here is as follows:

$$\min\Bigg[\sum_{i\in\mathcal{N}_{\mathrm{P}}}\sum_{j\in\mathcal{N}_{\mathrm{D}}}\sum_{t\in\mathcal{T}} C_{\mathrm{F}ij}x_{ijt} + \sum_{i\in\mathcal{N}_{\mathrm{P}}}\sum_{j\in\mathcal{N}_{\mathrm{D}}}\sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}} C_{\mathrm{V}ijk}q_{ijkt} + \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}} C_{\mathrm{L}ik}s_{\mathrm{L}ikt} + \sum_{i\in\mathcal{N}}\sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}} C_{\mathrm{U}ik}s_{\mathrm{U}ikt}\Bigg] \tag{4.101}$$

subject to

$$\sum_{k\in\mathcal{K}} q_{ijkt} - U_{ij}x_{ijt} \leqslant 0, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, j \in \mathcal{N}_{\mathrm{D}}, t \in \mathcal{T}, \tag{4.102}$$

$$s_{ikt} - s_{ik(t-1)} + \sum_{j\in\mathcal{N}_{\mathrm{D}}} q_{ijkt} = R_{ikt}, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, k \in \mathcal{K}, t \in \mathcal{T}, \tag{4.103}$$

$$s_{jkt} - s_{jk(t-1)} - \sum_{i\in\mathcal{N}_{\mathrm{P}}} q_{ijk(t-T_{ij})} = R_{ikt}, \quad \forall j \in \mathcal{N}_{\mathrm{D}}, k \in \mathcal{K}, t \in \mathcal{T}, \tag{4.104}$$

$$s_{ikt} + s_{\mathrm{L}ikt} \geqslant S_{\mathrm{SL}ik}, \quad \forall i \in \mathcal{N}, k \in \mathcal{K}, t \in \mathcal{T}, \tag{4.105}$$

$$s_{ikt} - s_{\mathrm{U}ikt} \leqslant S_{\mathrm{SU}ik}, \quad \forall i \in \mathcal{N}, k \in \mathcal{K}, t \in \mathcal{T}, \tag{4.106}$$

$$s_{ikt} \leqslant S_{\mathrm{MX}ik}, \quad \forall i \in \mathcal{N}, k \in \mathcal{K}, t \in \mathcal{T}, \tag{4.107}$$

$$q_{ijkt} \geqslant 0, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, j \in \mathcal{N}_{\mathrm{D}}, k \in \mathcal{K}, t \in \mathcal{T}, \tag{4.108}$$

$$x_{ijt} \in \{0, 1\}, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, j \in \mathcal{N}_{\mathrm{D}}, t \in \mathcal{T}, \tag{4.109}$$

$$s_{ikt}, s_{\mathrm{L}ikt}, s_{\mathrm{U}ikt} \geqslant 0, \quad \forall i \in \mathcal{N}, k \in \mathcal{K}, t \in \mathcal{T}. \tag{4.110}$$

The objective (4.101) minimizes the sum of the transportation and penalty costs. Constraints (4.102) together with the binary specifications in (4.109) force the ship usage variables to be equal to one for the ships in operation, so that we get the full setup cost for the ship voyages. Constraints (4.103) and (4.104) are the inventory balances at the loading and unloading ports, respectively, while constraints (4.105) and (4.106) calculate the safety stock shortfall and excess in the ports. The inventory starting level $S_{\mathrm{ST}ik}$ is used for s_{ik0} in (4.103) and (4.104). The upper storage capacity constraints at the ports are given in (4.107). Finally, the formulation involves binary requirements (4.109) and nonnegativity requirements (4.108) and (4.110).

This model (4.101)–(4.110) is reasonably easy to understand, but it is hard to solve because the solution of the linear relaxation will often transport small quantities to avoid penalty costs and just take the "needed" portion of the fixed sailing costs. Normally we will have $C_{\mathrm{L}ik} > C_{\mathrm{U}ik}$ in unloading ports and the other way around in loading ports.

Ronen (2002) used a model very similar to (4.101)–(4.110) to plan distribution from refineries. He presented the model without any upper safety

stock penalties but mentioned the use of such penalties in the discussion. Constraints (4.105) were given as equality constraints with explicit variables for lower safety stock excess. We get the same variables as surplus variables. Formulation (4.101)–(4.110) should make the LP marginally faster to solve. Ronen (2002) used CPLEX 6.5 to solve his model. For very small cases CPLEX managed to find the optimal solution with a user chosen relative tolerance of 1% for cutting off nodes in the branch-and-bound tree.

To be able to solve larger problems, we need to generate some cuts that restrict the number of x_{ijt} variables that can be 1, so that many x_{ijt} variables will be fixed to 0 after fixing some to 1.

Ronen (2002) added the following constraints to the model (4.101)–(4.110):

$$\sum_{k \in \mathcal{K}} q_{ijkt} - x_{ijt} \geqslant 0, \quad \forall i \in \mathcal{N}_{\mathrm{P}}, j \in \mathcal{N}_{\mathrm{D}}, t \in \mathcal{T}. \tag{4.111}$$

If we look at this as a valid cut, it is usually far from sharp enough. But if the q_{ijkt} variables are scaled such that the ship capacities, U_{ij}, have values slightly greater than 1, then constraints (4.111) will force all ships branched to be used to be nearly full. If the cost structure is such that we know that it is optimal to have all ships nearly full, then we can use (4.111) with scaled q_{ijkt} variables or better with a constant slightly less than U_{ij} in front of the x_{ijt} variable. This might make the problem easier to solve.

In addition to solving the model by use of commercial optimization software for smaller sized problems, Ronen (2002) presented a cost-based heuristic algorithm to assure that acceptable solutions were obtained quickly.

4.3.3 Other maritime supply chain applications

Reported research of more complex maritime supply chains is scarce. However, we will briefly refer to some existing studies.

A tactical transshipment problem, where coal is transported at sea from several supply sources to a port with inventory constraints was studied by Shih (1997). The coal is then transported from the port to several coal fired power plants. The objective is to minimize the procurement costs, transportation costs, and holding costs. Constraints on the system include company procurement policy, power plant demand, port unloading capacity, blending requirements, and safety stocks. The study proposes a MIP model for a real problem faced by the Taiwan Power Company. Kao et al. (1993) present a similar problem for the same company. They applied inventory theory to determine an optimal shipping policy. The underlying inventory model is nonlinear where the procurement costs, holding costs, and shortage costs are minimized subject to inventory capacity constraints. Liu and Sherali (2000) extended the problem described by Shih (1997), and included the coal blending process at the power plants in the mathematical model. They present a MIP model for finding optimal shipping and blending decisions on an annual basis. The solution procedure employs heuristic rules in conjunction with a branch-and-bound algorithm.

In a supply chain for oil, several types of models dealing with logistics are necessary. Chajakis (1997) describes three such models:

(a) crude supply – models for generating optimal short-term marine-based crude supply schedules using MIP,

(b) tanker lightering – models of tanker lightering, which is necessary in ports where draft or environmental restrictions prevent some fully loaded vessels from approaching the refinery unloading docks. Both simulation and MIP based tools are used, and

(c) petroleum products distribution – a simulation model that was developed for analyzing products distribution by sea.

However, several legs of the supply chain are not included in these models. In Chajakis (2000) additional models for freight rate forecasting, fleet size and mix, and crew planning are discussed.

A planning model for shipments of petroleum products from refineries to depots and its solution method is described by Persson and Göthe-Lundgren (2005). In the oil refining industry, companies need to have a high utilization of production, storage and transportation resources to be competitive. Therefore, the underlying mathematical model integrates both the shipment planning and the production scheduling at the refineries. The solution method is based on column generation, valid inequalities and constraint branching.

4.4 Fleet deployment in liner shipping

Liner shipping differs significantly from the other two types of shipping operations, industrial and tramp shipping, discussed so far in Section 4. However, also liner shipping involves decisions at different planning levels. Strategic planning issues were discussed in Section 3.2 for liner fleet size and mix and in Section 3.3 for liner network design. The differences among the types of shipping operations are also manifested when it comes to routing and scheduling issues. One main issue for liners on the tactical planning level is the assignment of vessels to established routes or lines and is called *fleet deployment*.

As discussed in Section 1, during the last four decades general cargo has been containerized and we have evidenced a tremendous increase in container shipping. This increased number of containerships almost always falls in the realm of liner shipping. Despite this fast growth, studies on deployment in liner shipping are scarce.

In this section, we want to focus on a fleet deployment problem where we utilize the different cruising speeds of the ships in the fleet. The routes are predefined, and each route will be sailed by one or more ships several times during the planning horizon. Each route has a defined common starting and ending port. A round-trip along the route from the starting port is called a voyage.

The demand is given as a required number of voyages on each route without any explicit reference to the quantities shipped. The fleet of ships is heterogeneous, so we can reference quantities implicitly by saying that not all ships

can sail all routes. Such a specification can incorporate needed ship capacity together with compatibilities between ships and ports. With information about the feasible ship-route combinations, we do not need to keep track of the loads on board the ships. Further, the routes do not need to share a common hub. Figure 12(a) presents a case with one hub, and Figure 12(b) presents one with several hubs.

The ships in the available fleet have different cruising speeds. Each ship is assigned to a single route and is not allowed to switch routes during the planning horizon. The fleet deployment problem consists of determining which route each ship is going to sail. The planning goal is to minimize the cost of the ships.

In the mathematical description of the problem each ship type is represented by an index v and the set of ship types is given by $\mathcal{V}$. Let $\mathcal{R}$ be the set of routes and $\mathcal{R}_v$ the set of routes that can be sailed by a ship of type v. The elements in both sets are indexed by r.

V_v is the number of ships available of type v. The number of voyages during the planning horizon along route r for a ship of type v is represented by $N_{\mathrm{VY}vr}$. Normally this is not treated as an integer number of voyages. The demand is given by $D_{\mathrm{V}r}$ which is the required minimal number of voyages along route r during the planning horizon. T is the length of the planning horizon in days, and is one year for the underlying real problem. S_{vr} represents the shipping season for a ship of type v operating on route r. The shipping season S_{vr} is the total length of the service periods for ship type v during the planning horizon. This means that if a ship is allocated to a route, it is operating on that route during its whole shipping season.

Often, the demand requirement is such that we, for instance, are allowed to combine 3.7 voyages of one ship with 8.4 voyages of another ship to get a total of 12.1 voyages to meet a demand of 12 voyages. In such cases, it is not necessary for $N_{\mathrm{VY}vr}$ to be integer. This also gives S_{vr} equal to the time a ship of type v is available during a year independently of route r.

However, if we want to be sure that each port on route r is visited at least $D_{\mathrm{V}r}$ times during the planning horizon, we need to calculate $N_{\mathrm{VY}vr}$ as an integer. Then S_{vr} is calculated as the number of whole voyages multiplied by the time of one voyage. This is the reason for defining the shipping season for a ship dependent on the route.

The cost consists of two parts. First, the cost of operating a ship of type v on route r during the planning horizon is given by $C_{\mathrm{Y}vr}$. Secondly, we have the lay-up cost. The days the ship is out of service for maintenance or other reasons are called lay-up days. The cost for each lay-up day for a ship of type v is denoted by $C_{\mathrm{E}v}$.

To make the model similar to the models in the literature, we use the following types of decision variables: the fleet deployment variables, s_{vr}, $v \in \mathcal{V}$, $r \in \mathcal{R}_v$, represents the integer number of ships of type v allocated to route r, and d_v, $v \in \mathcal{V}$, gives the total number of lay-up days for ships of type v.

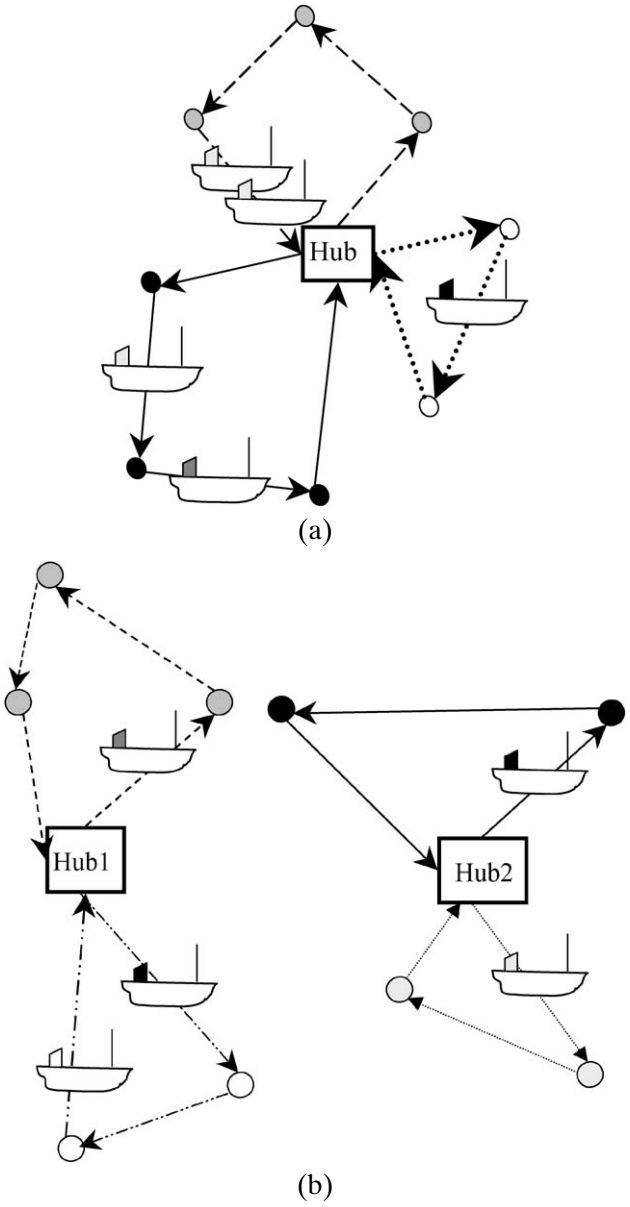


Fig. 12. (a) Fleet deployment with nonoverlapping routes and a common hub. (b) Fleet deployment with nonoverlapping routes and several hubs.

The mathematical formulation of this fleet deployment problem for ships with different operating speeds and capacities is as follows:

$$\min\left[\sum_{v\in\mathcal{V}}\sum_{r\in\mathcal{R}_v} C_{\mathrm{Y}vr}s_{vr} + \sum_{v\in\mathcal{V}} C_{\mathrm{E}v}d_v\right] \tag{4.112}$$

subject to

$$\sum_{r \in \mathcal{R}_v} s_{vr} \leqslant V_v, \quad \forall v \in \mathcal{V}, \tag{4.113}$$

$$\sum_{v \in \mathcal{V}} N_{\mathrm{VY}vr} s_{vr} \geqslant D_{\mathrm{V}r}, \quad \forall r \in \mathcal{R}, \tag{4.114}$$

$$d_v + \sum_{r \in \mathcal{R}_v} S_{vr} s_{vr} = V_v T, \quad \forall v \in \mathcal{V}, \tag{4.115}$$

$$s_{vr} \geqslant 0 \text{ and integer}, \quad \forall v \in \mathcal{V}, r \in \mathcal{R}_v, \tag{4.116}$$

$$d_v \geqslant 0, \quad \forall v \in \mathcal{V}. \tag{4.117}$$

Here (4.112) is the total cost of sailing the routes with the selected ships and the cost of the lay-up days. Constraints (4.113) prevent the number of ships in operation from exceeding the number available, while constraints (4.114) ensure that each route is sailed at least the required number of voyages demanded. The lay-up days for each ship type are calculated in constraints (4.115). Finally, the formulation involves integer and nonnegativity requirements on the fleet deployment variables and lay-up variables, respectively.

Powell and Perakis (1997) presented this model using a different notation. The model was tested on a real liner shipping problem and substantial savings were reported compared to the actual deployment. Powell and Perakis (1997) used standard commercial software for the formulation (AMPL) and solution (OSL) of their model. The example they give has 11 types of ships and 7 routes with an average number of required voyages just below 20. All their data for the number of voyages for ships of a given type on a given route are noninteger.

We have assumed here that a ship allocated to a route is just operating on that route during its whole shipping season, even if that results in more voyages than required on that route. This means that the model does not allow for a choice between extra voyages or extra lay-up days.

Constraints (4.115) calculate the total number of lay-up days for each ship type. It is reasonably easy to remove these constraints from the model by a reformulation. Since each ship is used only on one route, we can pre-calculate the number of lay-up days for a ship of type v that is used on route r, before the optimization and add the corresponding lay-up cost to the annual cost of using the ship on that route. This also removes variable d_v. If we want an integer number of voyages for each ship, we need to divide the lay-up days calculated by (4.115) into two parts, one part for the real lay-up days for maintenance, and one part where the ship only waits for the next planning horizon. The cost per day for each of these parts may be different, and this difference is most easily taken care of in a pre-calculation phase.

The formulation (4.112)–(4.117) is a tactical planning model. If we want to use it in a pure strategic planning situation, we will normally assume that we can buy or build as many ships as we want of each type. Then constraints (4.113) will not be binding and the optimization problem decomposes into a

problem with one constraint, (4.114), for each route after pre-calculating the lay-up cost and removing (4.115).

The work presented by Powell and Perakis (1997) is an extension and improvement of the work in Perakis and Jaramillo (1991) and Jaramillo and Perakis (1991). In the latter two papers, an LP approach is used to solve the fleet deployment problem. Manipulation of the results is needed to achieve integer solutions from the continuous ones, which may lead to a suboptimal solution and even violation of some constraints.

Section 3.2.2 discussed a strategic fleet size and mix model originally given by Fagerholt and Lindstad (2000). With a fixed fleet that model becomes a tactical fleet deployment model.

Earlier fleet deployment studies for assigning vessels to origin–destination port pairs can be found in Papadakis and Perakis (1989), Perakis and Papadakis (1987a, 1987b) and Perakis (1985). Various models were presented where both full and ballast speeds and several additional constraints were considered.

4.5 Barge scheduling

Barges usually operate in protected bodies of water, generally in inland waterways. Barges can be self-propelled or they may be towed by a tugboat, or pushed by a tugboat. Often multiple barges are combined into a single tow that is pushed or pulled by a single tug. On the Mississippi River system a barge can load up to 600 tons and a tow is composed of up to 15 barges. Since barges operate on inland waterways they must follow the navigable waterway and therefore their routes are linear like coastal routes or, if there are branches in the waterway, the routes may have a tree shape. Loaded and/or empty barges can be added to a tow or dropped off from a tow in ports that are passed-by along the route of the tow. Barges often have to pass through locks on their way up or down the river. This complicates their scheduling because they may have to wait for their turn to pass through a lock, and locks may have limited hours of operation. Research on barge transportation is scarce. Several papers discussing fleet design were discussed in Section 3.

Scheduling of barges in inland waterways is similar to scheduling ships with the additional complications that may be posed by locks. Such additional constraints may be very important in barge scheduling, but may be hard to incorporate in scheduling models similar to those described in Section 4.1.

Very few works are dedicated to barge scheduling. The initial work on scheduling barges was provided by O'Brien and Crane (1959) who used simulation to evaluate the impact of various scheduling policies on fleet size and mix requirements. Schwartz (1968) suggested a linear MIP model for scheduling a fleet of tugs and barges for the delivery of a given set of cargoes at minimal cost. The size of the model was far beyond the capabilities of solution algorithms at that time. A special barge scheduling problem that involves high uncertainty in timing of activities was discussed by Vukadinovic and Teodorovic (1994) and later by Vukadinovic et al. (1997). The barges are used to move gravel from a

dredging site and are moved in tows by pushing tugs. The barge loading and discharging process is subject to significant uncertainty regarding its timing. The key decision is the assignment of loaded barges to tugs for a planning horizon of one day. There is a single loading location and multiple discharging ports, but a loaded barge is unloaded in a single port. The initial paper used fuzzy logic to suggest a schedule, and the second one proposed a neural network that learns from examples and can emulate the dispatcher's decision making process.

4.6 Scheduling naval vessels

In contrast to commercial vessels that are usually used to transport one type of cargo or another, the main mission of naval and coast guard vessels is to perform assigned tasks at sea. Such tasks may include patrols, training, exercises, law enforcement, search and rescue, and others. In smaller navies, naval vessels usually stay close to home and return to base frequently. However, in larger navies, naval vessels may spend extended periods of time at sea, and have to be resupplied at sea. Naval vessels also spend lengthy periods of time at port or yards for maintenance, renovation, and training. Usually the major objective in scheduling naval and coast guard vessels is to assign the available fleet to a set of specified tasks in a manner that maximizes or minimizes a set of measures of effectiveness. First we discuss scheduling naval combatants, then we move to scheduling coast guard vessels, and we close with logistical support at sea.

4.6.1 Scheduling naval combat vessels

Scheduling an available naval fleet to perform a planned set of activities is a classical naval application. Such activities may include major operations, exercises, maintenance periods, and other events. Brown et al. (1990) considered the problem of determining the annual schedule of the US Navy's Atlantic Fleet combatants. The goal is to assign ships to events in a manner that meets all the event requirements and minimizes deviations from ideal schedules for individual ships. Each event requires a given number of units of particular vessel types and weapon systems. A generalized set partitioning model is used to solve the problem optimally. Intricate realistic schedule constraints can be incorporated in the schedule generator.

The same problem is addressed by Nulty and Ratliff (1991), but in a somewhat different manner. An integer programming formulation is developed, but results in a model of prohibitive size. This fact combined with the qualitative nature of additional secondary objectives and constraints suggest an interactive optimization approach. The proposed approach allows the user to generate a good initial fleet schedule by using network algorithms, and improve the solution by interactively addressing the issues that are difficult to quantify. They also suggest that the method of Brown et al. (1990) could be used to find a starting solution for the interactive procedure.

4.6.2 Scheduling coast guard vessels

A problem that is essentially similar to scheduling naval combatants is faced by coast guards. However, coast guard vessels stay closer to their home base and generally do not have to be resupplied at sea. A typical problem is to schedule a fleet of coast guard cutters (vessels) to perform a set of assignments. Each assignment has a given duration, and a desired number of cutters. Such a problem was addressed by Darby-Dowman et al. (1995). In their model the requirements are treated as goals, and not meeting a goal is allowed but penalized. The problem is solved by using a set partitioning model. The objective is to select the set of schedules that provides a solution that is as close to meeting the requirements as possible. The system was originally intended for use in determining operational schedules. However, additional use to address strategic issues such as future operating policy and fleet mix arose during the project.

A system for solving similar scheduling problems for the US Coast Guard cutters was presented by Brown et al. (1996). They developed costs and penalties for the model to mimic the motives and rules of thumb of a good scheduler. The objective was to minimize the costs, and the elastic MIP model was solved on a personal computer within a few minutes.

Another type of vessel scheduling problem faced by a coast guard is routing and scheduling buoy tenders. These vessels are used to service and maintain navigational buoys. Cline et al. (1992) describe a heuristic algorithm for routing and scheduling US Coast Guard buoy tenders. Each buoy has a service time window dictated by the planned maintenance schedule. Since each tender has the sole responsibility for servicing its set of buoys, the problem is decomposed into a set of traveling salesman problems with time windows, one for each tender. They used a best-schedule heuristic to solve the problem.

4.6.3 Logistical support at sea

Supporting naval vessels at sea poses additional challenges. Schrady and Wadsworth (1991) described a logistic support system that was designed to track and predict the use of consumable logistic assets (fuel, ordnance) by a battle group. The system was tested during fleet exercises and was quite successful. Williams (1992) dealt with the replenishment of vessels at sea. He presented a heuristic algorithm to replenish a group of warships at sea while the ships carry out their assignments. The heuristic rules were derived from replenishment experts and are based on experimentation.

4.7 Ship management

Several topics fall into the category of ship management and we shall discuss briefly the following ones: crew scheduling, maintenance scheduling, positioning of spare parts, and bunkering. Deep-sea vessels spend extended periods of time at sea and the crew lives on board the ship. Short-sea vessels make frequent port calls and the crew may live on shore. This difference has significant

impact on ship management issues. Crew scheduling for deep-sea vessels is not a major issue. Crew members spend months on the vessel and then get a long shore leave. For short-sea vessels the crew may change frequently, and crew scheduling may be an issue. A special type of such crew scheduling problem is presented by Wermus and Pope (1994).

Numerous mechanical and electrical systems are installed on board a ship and they require maintenance. A ship is usually scheduled once a year for maintenance in a port or a shipyard, and once every several years a ship is surveyed by its classification society in a shipyard. However, some maintenance is required between such planned maintenance periods, and this includes both routine/preventive maintenance, and repair of breakdowns, at least temporarily, till the ship reaches the next port. On-board maintenance is usually done by the crew, but the shrinking size of crews reduces the availability of the crew for maintenance work. A large ship may have less than two dozen seamen on board, and that includes the captain and the radio officer. This limited crew operates the ship around the clock. A specialized analysis of repair and replacement of marine diesel engines was presented by Perakis and Inozu (1991). In order to facilitate maintenance a ship must carry spare parts on board. The amount of spare parts is determined by the frequency of port calls and whether spares and equipment are available in these ports. Large and expensive spares that cannot be shipped by air, such as a propeller, may pose a special problem, and may have to be prepositioned at a port or carried on board the vessel.

A ship may consume tens of tons of bunker fuel per day at sea, and there may be significant differences in the cost of bunker fuel among bunkering ports. Thus one has to decide where to buy bunker fuel. Sometimes it may be worthwhile to divert the ship to enter a port just for loading bunker fuel. The additional cost of the ship's time has to be traded off with the savings in the cost of the fuel.

5 Operational planning

When the uncertainty in the operational environment is high and the situation is dynamic, or when decisions have only short-term impact, one resorts to short-term operational planning. In this section we discuss operational scheduling where only a single voyage is assigned to a vessel, environmental routing where decisions are made concerning the next leg of the voyage, speed selection, ship loading, and booking of single orders.

5.1 Operational scheduling

The demarcation between tactical and operational scheduling in industrial and tramp shipping is fuzzy, and therefore Section 4.1 considered both planning levels. However, there are some situations that can be placed solely on

the operational planning level and they are discussed here. In certain circumstances it is impractical to schedule ships beyond a single voyage. This happens when there is significant uncertainty in the supply of the product to be shipped, or in the demand for the product in the destination markets. The shipped product may be seasonal and its demand and supply may be affected by the weather. These factors contribute to the uncertainty in the shipping schedule. Citrus fruit is an example of such a product. This is a highly seasonal product that is shipped in large quantities over several months a year, and may require refrigerated vessels. The shipper has to assure sufficient shipping capacity in advance of the shipping season, but does not know in advance the exact timing, quantities, and destinations of the shipments. The shipper, who owns the cargo, does not have return cargoes for the ships, so the ships are hired under contracts of affreightment or spot charters, and generally do not return to load a second voyage. Thus every week the shipper has ships available for loading in the producing area and either a load is assigned to each ship or demurrage is paid for the ship. Based on product availability, demand projections, inventory at the markets, and transit times, the shipper builds a shipping plan for the upcoming week, and has to assign the planned shipments to the available fleet at minimal cost. Usually the contract of most vessels hired for a single voyage confines them to a range of unloading ports. In some operations a vessel may unload in more than one port, and the requirement of a destination port may exceed the size of the largest vessel and can be split among several vessels.

Ronen (1986) discussed such an operational scheduling problem, presented a model and a solution algorithm that provided optimal solutions to smaller size problems, and heuristics for solving larger problem instances. Later Cho and Perakis (2001) suggested a more efficient formulation to the same problem that is a generalized version of the capacitated facility location problem.

5.2 Environmental routing

Ships navigate in bodies of water and are exposed to a variety of environmental conditions, such as: currents, tides, waves, and winds. Recognizing these conditions is the first step toward selecting routes that mitigate their effects, or even take advantage of them. Generally, when a ship moves between two ports it has to select its route through the body of water. However, such a choice is very limited in coastal and inland waterway navigation. Proper selection of the route may assure on-time arrival at the destination port, or even shorten the time of the passage and reduce its cost. The terms *environmental routing* and *weather routing* are often used interchangeably although the second one is a subset of the first. Weather is part of the environment in which ships operate, and it affects the waves encountered by ships. We confine our short discussion to the impact of waves and ocean currents. Material concerning tides and winds can be found in the naval architecture, navigation, and meteorology literature.

5.2.1 Waves

Waves may have a significant impact on route selection. In order to take waves into account one has first to know their height and direction along the contemplated route as a function of location (x and y coordinates). Such knowledge may allow selection of the route and of power setting that minimize the transit time. However, the waves' height and direction may change over time, and may not be known in advance. Papadakis and Perakis (1990) analyzed a minimal time vessel routing problem under stationary conditions that is appropriate for relatively short passages. Given wave height and wave direction as a function of location, select the route and power setting of the vessel that minimize the transit time. Local optimality considerations combined with global boundary conditions resulted in piecewise continuous optimal policies. They used variational calculus and optimal control theory in their analysis. Perakis and Papadakis (1989) extended their analysis of the minimal time vessel routing problem to a time-dependent environment, where the sea condition at any point changes over time. This allows analysis of longer passages. In addition they considered voyages consisting of multiple legs with port calls of known length between the legs. Although they provide a numerical example, no estimates of potential benefits (or savings) are available. However, they show that when the objective is to minimize transit time "wait for a storm to pass" policy is never optimal. Instead "one should go ahead under the maximum permissible power setting. . .".

5.2.2 Ocean currents

In most oceans there are regular currents that ships may be able to exploit for faster passage. Lo et al. (1991) estimated that by exploiting ocean currents the world commercial fleet could reduce its annual fuel costs by at least $70 million. They also provide anecdotal evidence that some operators try to take advantage of prevailing currents. However, this is easier said than done. Ocean currents are not constant but rather change over time. Thus getting reliable timely information regarding the ocean current at the location of a vessel poses a major obstacle. Satellite altimetry may provide timely reliable estimates of dynamic current patterns that are necessary for routing a vessel through such currents. McCord et al. (1999) took a closer look at potential benefits of such data. Their study uses dynamic programming to optimize ship routes through estimated current patterns in a dynamic area of the Gulf Stream. They conclude that elimination of data bias and present sampling limitations can produce about 11% fuel savings for a 16-knot vessel. They found that the contribution of such routing is much better on with-current voyages than on counter-current voyages. The major question is whether there is a sufficient market to justify development of a system for collection of the necessary data.

Environmental routing is complicated by the complexity of the continuous dynamic environment in which it takes place, and the lack of the necessary

timely reliable data. Due to these reasons environmental routing seems to be in its infancy and is a fertile ground for further research.

5.3 Speed selection

A ship can operate at a speed slower than its design speed and thus significantly reduce its operating cost. However, a ship must maintain a minimal speed to assure proper steerage. For most cargo vessels the bunker fuel consumption per time unit is approximately proportional to the third power of the speed (the consumption per distance unit is proportional to the second power of the speed). Thus, reducing the speed by 20% reduces the fuel consumption (per time unit) by about 50% (or by about 36% for a given sailing distance). When bunker fuel prices are high the cost of bunker fuel may exceed all other operating costs of the ship. Thus there may be a strong incentive to steam at slower speed and reduce the operating costs. In the wake of the high fuel price during the 1970s, Ronen (1982) presented three models for the determination of short-run optimal speed for different types of legs:

- an income generating leg,
- a positioning (empty/ballast) leg, and
- a leg where the income depends on the speed.

When one widens the horizon beyond a single vessel, the perspective may change. A fleet operator that controls excess capacity can reduce the speed of the vessels and thus reduce the effective capacity of the fleet, instead of laying-up, chartering-out or selling vessels.

Under various operational circumstances a scheduler has to assign an available fleet of vessels to carry a specified set of cargoes among various ports. Often cruising speed decisions may be an inherent part of such fleet scheduling decisions. Cruising speed decisions affect both the effective capacity of the fleet and its operating costs.

Under a contract of affreightment (COA) a ship operator commits to carry specified amounts of cargo between specified loading port(s) and unloading port(s) at a specific rate over a specific period of time for an agreed upon revenue per delivered unit of cargo. The term *fleet deployment* is usually used for ship scheduling problems associated with liners and with COAs, because the vessels are essentially assigned to routes that they follow repeatedly, and the deployment decisions cover longer terms. Perakis and Papadakis (1987a, 1987b) determined the fleet deployment and the associated optimal speed, both loaded and in ballast, for ships operating under a COA between a single loading port and a single unloading port. A more comprehensive version of this problem was later dealt with by Papadakis and Perakis (1989). They expanded the problem to address multiple loading ports and multiple unloading ports, but still assumed that each ship returns in ballast to its loading port after unloading its cargo. They used nonlinear programming to determine the vessel allocation to the routes and their cruising speed, both loaded and in ballast.

Tramp and industrial operators usually face shorter term ship scheduling problems. A set of cargoes has to be carried by the available fleet, and if the fleet has insufficient capacity some cargoes may be contracted out. The cruising speed of the vessels in the available fleet can be an inherent part of the scheduling decisions. Bausch et al. (1998) and Brown et al. (1987) addressed this situation, and in their work the cruising speed was determined simultaneously with the schedule. Whereas the last two papers had hard time windows for loading and unloading the cargoes, Fagerholt (2001) considered also soft time windows, a situation that allows more flexibility in determining the cruising speed of the vessels, and may result in a lower cost schedule.

In addition to cost and schedules, short-term cruising speed decisions should take into account also the impact of the destination port operating times. If the destination port is closed over the weekend (or at night) there is no point arriving there before the port opens. Thus reducing the cruising speed and saving fuel makes sense. In the case where cargo-handling operations of a vessel that started when the port was open continue until the vessel is finished, even after the port closes, it may be worthwhile to speed up and arrive at the destination port to start operations before it closes. A more detailed discussion of these tactics is provided in Section 6.2.

5.4 Ship loading

A ship must be loaded in a safe manner in order to prevent loss of the ship or damage to the cargo. Ships are designed with certain types of cargo in mind. A crude tanker is designed to carry crude oil, and a containership is designed to carry containers. A ship floats on water and its stability must be assured during passage as well as in port. Ballast tanks are built into the hull of a ship in order to help maintain its stability by filling them with seawater. When a ship is full with cargo of a uniform density for which it is designed, such as crude oil or iron ore, usually there are no stability problems. Stability problems arise when (a) a ship is partially loaded, then the weight distribution of the cargo must be properly planned and monitored, both while sailing at sea and during loading or unloading operations in port, or (b) the cargo is not properly secured and may shift during passage, for example, liquid bulk cargo may slosh in partially empty tanks, or (c) when the ship is fully loaded with nonuniform cargo, such as containers or general cargo. In such a case an improper weight distribution of the cargo may result in excessive rolling or pitching that may lead to loss of the ship. In extreme cases improper weight distribution may cause excessive structural stress that may lead to break up of the ship.

Ship stability has several dimensions. The *Trim* of a ship is the difference between the forward and aft draft, and must remain within a narrow range. There also should be balance between weight of the cargo on the port (left) side and the starboard (right) side of the ship so it will remain horizontal. The center of gravity of the ship should not be too high in order not to make the

ship "top heavy" and easy to roll, and not too low so the ship will not snap back too fast from a roll which may cause on-deck cargo to break loose.

The more complex ship loading problems are encountered in loading containerships. Not only the stability of the vessel has to be assured but also the efficiency of cargo handling operations in the current and following ports must be taken into account. Containers have different weights and that may affect the vessel stability. Due to the design of containerships access to a specific container may be obstructed by other containers stowed on top of it. Thus container shifting may be necessary to unload a specific container. Therefore, in order to minimize future container shifting operations one has to take into account the destination port of the loaded containers when one decides where to load them onboard the vessel. Moreover, one also has to consider the destination ports of the containers that will be loaded in following ports of call, and some of these containers may not even be booked yet. There may also be containers stuffed with dangerous goods. Such containers impose additional constraints due to spatial separation requirements.

The focus of research on ship loading has been on loading container ships. A good description of the various considerations involved in containership loading is provided by Martin et al. (1988). They developed heuristics that emulate strategies used in manual load planning and showed some improvements in materials handling measures.

Avriel et al. (1998) focused on minimizing container shifting. They formulated a binary linear program for the container stowage planning problem that minimizes the number of container shifting operations. Since the problem is NP-complete they designed a "suspensory heuristic" to achieve a stowage plan. Their work is of limited applicability because it assumes away stability and strength requirements, accommodates only one size of containers, and ignores hatch covers.

A comprehensive approach for planning container stowage on board containerships is provided by Wilson and Roach (2000). Their objective is to find a stowage plan that assures that no ship stability or stress constraints are violated, and minimizes container shifts (re-handles). Additional considerations are reduction of the ballast required by the vessel and efficient use of cranes when loading and unloading. Wilson and Roach described a computerized methodology for generating commercially viable stowage plans. All characteristics of the problem are considered, but optimality is not necessarily sought. Their stowage planning process is broken down into two phases, (a) "strategic planning" where "generalized" containers are assigned to "blocks" of cargo space, and (b) "tactical planning" where specific containers are assigned to specific slots within the blocks determined earlier. This approach significantly reduces the combinatorial complexity of the problem. Their objective consists of a dozen different criteria that are assigned weights. The strategic planning phase uses a branch-and-bound search, and the tactical planning phase uses a tabu search. They tested their methodology on commercial data for a 688 TEU vessel with a mix of container sizes and types, and four destination ports. Com-

mercially viable solutions were received in a couple of hours on a 166 MHz computer. These solutions were comparable with those generated by experienced human planners. However, it takes a human planner several days to get such solutions.

A similar two-stage approach is used by Kang and Kim (2002) to generate container stowage plans. In the first stage they assign containers to holds for each port separately by solving a problem similar to a fixed charge transportation problem using a heuristic based on the transportation method. In the second stage they assign containers to slots for each hold separately using a tree search procedure. Since the first stage is done for one port at a time the resulting stowage plan may be problematic. Therefore they iterate between the two stages to improve the plan. They tested their approach on randomly generated problems and compared their results to a couple of earlier suggested models. However, they admit the limited applicability of their approach because it considers only one size of containers (40′), and does not consider refrigerated containers or ones with hazardous materials.

The container stowage planning problem is very complex and we are far from finding optimal solutions, or even agreeing on the components of the objective function. The related problem of stowage sequencing, which represents the port's perspective, is discussed at length in the chapter by Crainic and Kim (2007).

5.5 Booking of single orders

An important operational problem in commercial shipping is booking of single orders. Since a shipper expects an acceptance/rejection decision on a single cargo request more or less immediately, for the shipping company the problem consists of deciding whether to accept a single cargo or not. This problem is somewhat different between liner and tramp/industrial shipping. In liner shipping, where a single cargo is usually a small fraction of the vessel's capacity, it is usual to accept a cargo if there is space available on the given ship line, and to reject or suggest another time of departure if not. However, sometimes it may not be profitable to accept a cargo even if there is space available, as there may appear requests for better paying cargoes later on. This problem of stochastic optimization in liner shipping is rarely dealt with in the literature. The authors are aware of only a single reference on the subject, and it is a rather out-dated conference contribution (Almogy and Levin, 1970).

In tramp shipping it is also usual to accept a single cargo request if the planner is able to find space available. To see if there is space available, rescheduling the whole fleet with the existing cargo commitments together with the new optional cargo may be necessary because a single cargo may take a large share of a vessel's capacity, or even be a full shipload. This is thoroughly discussed in Section 4.1.5. Industrial shipping is similar in this respect to tramp. However, also in tramp shipping, as for liner shipping, it may sometimes be advantageous

not to accept a single cargo request as more profitable cargoes may appear later. The authors are not aware of any published work on this aspect.

6 Robustness in maritime transportation

As discussed in previous sections, there are many uncertain factors in the ocean shipping industry resulting in delays and lack of timely fulfillment of plans. Therefore, in order to encourage trust in the planning process, it may often be important to consider robustness in optimization models used for planning. Despite this, models that have been developed for the shipping industry only rarely deal with these aspects.

In this section we discuss a few problems from the shipping industry where uncertainty and robustness play important roles, as well as approaches for achieving more robust solutions. It should be emphasized that this section does not present a comprehensive overview but rather provides several examples. In Section 6.1 we concentrate on strategic planning problems, while tactical and operational planning problems are considered in Sections 6.2 and 6.3, respectively. Section 6.4 discusses optimization and persistence.

6.1 Strategic planning and uncertainty

The most important strategic planning problem for all shipping segments (industrial, tramp, and liner) is probably fleet sizing and composition. However, the quality of decisions regarding this aspect is strongly influenced by many uncertainties, probably much more than decisions for any shorter planning horizon. There are several major reasons for this uncertainty:

- The long time horizon that these decisions span, which can be several years. In some cases, when the decision involves building new ships, it may span up to 20–30 years.
- Demand for shipping is a derived demand. It depends on the level of economic activity, prices of commodities, and other factors.
- There is a significant time lag between changes in demand for maritime transportation and the corresponding adjustments in the capacity of such services.

During such a long time horizon one will experience major unpredictable fluctuations both in the demand for shipping services and on the supply side. These factors are highly interwoven. For instance, if demand for transport services within a given market segment increases, we would probably see an increase in both freight rates and ship prices, and the same is true in the opposite direction.

Another important strategic decision that is relevant to all shipping segments is whether a shipping company should accept a long-term contract or not. In such a long-term contract, the shipping company is typically committed

to carry a specific quantity more or less evenly distributed over the contract period, and receives a given revenue per unit of cargo lifted. Also here, the decision should be made only after cautious consideration (or speculation) regarding the direction that the market will take in the future. If, for instance, the spot market experiences a boost and the actual freight rates increase it would be unfortunate to have most of the fleet tied up in contracts at lower rates. On the contrary, if the market dips, it would be advantageous to have a substantial contractual coverage, in order to ensure both income and engagement for the ships.

There are different approaches for handling uncertainty and robustness, such as:

- simulation,
- re-optimization for different scenarios or input parameters,
- adding slack to the input parameters (e.g., service speed),
- deterministic models that incorporate penalties, and
- stochastic optimization models.

Simulation is a simple approach that is used to consider stochastic conditions and uncertainty. There are some examples where simulation models have been used for strategic planning purposes in the shipping industry, see, for instance, Darzentas and Spyrou (1996), Richetta and Larson (1997), and Fagerholt and Rygh (2002).

Another simple approach for considering uncertainty is to make several runs with an optimization model for different scenarios. In this way, one can decide what is the optimal decision for a given scenario or for a given set of input parameters. The problem in using this method is that solutions are often not robust and are strongly affected by the specific set of values used for the input parameters. Since flexibility is not built into the plans, extreme solutions are often produced.

Stochastic conditions like the ones mentioned above and other ones can also be approached both by deterministic and stochastic optimization models. An example of using deterministic optimization models with penalties to achieve more robust solutions is discussed in the next section for a tactical ship scheduling problem. To the authors' best knowledge there are no published papers where stochastic optimization models are used for strategic planning in the shipping industry. The only one discussing the issue is by Jaikumar and Solomon (1987), where a model for determining the minimum number of tugs needed to move barges between ports on a river is presented. They discuss how their model can be extended to incorporate stochastic demands.

6.2 Robust tactical planning

In Section 4 we presented tactical planning problems and models for the different shipping segments. However, the models presented there and the solutions that can be obtained from them do not handle the uncertainty and

robustness aspects. Several unpredictable factors influence the fulfillment of plans and should often be considered in the planning process. The two most important are probably:

- weather conditions that can strongly influence the sailing time, and
- port conditions, such as strikes and mechanical problems that can affect the time in port.

A ship may often have to reduce its speed in bad weather. This may result in late arrival for the next planned cargo. In such cases the planner often has to reschedule the whole fleet. If the planner has built in enough slack in the schedule, the planned schedule may still be valid. However, since ships have high costs, very little slack is usually built into their schedules.

In some cases, ships may require high tide to get into the port fully loaded. In other cases empty barges may not be able to pass under bridges at high tide. In short-sea shipping applications where sailing times are short relatively to port times, and tides may have a significant impact on port access, a small delay may be amplified due to additional waiting for high tide. Many ports are also closed for cargo handling operations during nights and weekends. Cargo handling time that is longer than one working day of the port will span multiple days. This means that the ship will stay idle much of the time in port, and the total time in port will depend on the ship's arrival time.

Consider the following example. A ship has to load a cargo at a specified port. The loading time window starts on Wednesday at 8:00 and ends on the next Monday at 24:00. The operating hours of the port are between 8:00 and 16:00 from Monday to Friday. It takes 12 operating hours to load the cargo. Figure 13 shows the necessary time in port as a function of the arrival time of the vessel. We see that the total time spent in port varies from 28 to 92 hours, depending on the arrival time. Twenty eight hours is the minimal time spent in port, while 92 hours is the maximal time and includes a lot of idle time during the weekend.

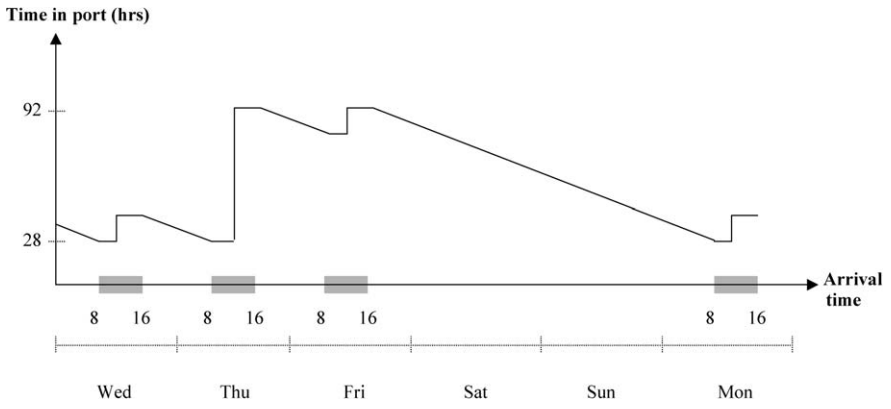


Fig. 13. Time spent in port as a function of arrival time.

A ship arriving on Wednesday morning at 8:00 will be loading for 8 hours on the first day and 4 hours on the next day. This gives a total of 28 hours in port. In the other extreme, take a ship arriving at 15:00 on Thursday. It loads for one hour on that day, stays idle for 16 hours during Thursday night and continues loading on Friday for 8 hours, but she does not finish loading before the port closes for the weekend. It has to continue loading on Monday morning at 8:00 and finishes at 11:00. This means that the ship stays idle in port for 64 hours during the weekend, giving a total of 92 hours in the port. It should be emphasized that in practice it may often be possible to negotiate a few hours extension to the loading/unloading operations, usually at a cost.

In these cases, a delay due to bad weather or port conditions may have even stronger effect than in other cases, as the delay may result in the ship staying idle in port during weekends. Christiansen and Fagerholt (2002) deal with such a problem. There, a deterministic solution method for making the schedules robust is presented. Their solution method is based on the set partitioning approach described in Section 4.2.2. However, to ensure schedules that are robust the concept of *risky arrival* is introduced. A risky arrival is defined as a planned arrival time in port that with only a moderate delay will result in the ship staying idle during a weekend. In order to reduce the number and magnitude of risky arrivals for a fleet schedule, Christiansen and Fagerholt (2002) calculate a penalty cost depending on how 'risky' the arrival time is. This penalty cost is calculated during the a priori schedule generation and is added to the other cost elements in the objective function in the set partitioning model. The computational results show that the planned fleet schedule's robustness can be significantly increased at the sacrifice of only small increases in transportation costs.

We can also find a few other contributions within ship scheduling where penalty costs are used in connection with time windows. In Fagerholt (2001), hard time windows are extended to soft ones by allowing late or early service, though at a penalty cost. Christiansen (1999) studies a combined ship routing and inventory management problem described in Section 4.3.1. The transported product is produced in some port factories and consumed in others. At all factories there are hard inventory limits for the transported product. In order to reduce the possibility of violating the inventory limits at the port factories Christiansen and Nygreen (2005) introduce an additional pair of soft inventory limits within the hard ones. Thus the soft inventory limits can be violated at a penalty, but it is not possible to exceed the stock capacity or to drop below the lower inventory limit. They show that the soft inventory constraints can be transformed into soft time windows.

Another problem regarding uncertainty and robustness in ship routing and scheduling is that in some cases the planner knows the loading port but the unloading port is not known at the time of loading. Sometimes just a geographical region is given for unloading, and the particular unloading port is specified after the voyage has started. In these cases the planner has several practical options:

- the planner can, based on his or her experience, make a qualified guess regarding the unloading port and use it for planning,
- use a port that is more or less in the middle of the specified unloading area as an "average",
- plan for worst-case, i.e., use the port that is farthest away in the area (e.g., farthest up in the river), and
- run different scenarios regarding the different optional unloading ports to see how the different alternatives affect the schedule.

6.3 Robust operational planning

Also operational problems in maritime transportation may pose robustness issues. Delays due to tides and restricted opening hours in ports, as discussed in the previous section, can often be regarded as operational ones. How to handle such delays when they occur is often referred to as "disruption management". Typically for shipping, it is often possible to increase the ship's speed to some extent when a delay occurs. However, this comes at the sacrifice of much higher fuel consumption, see Section 5.3 on speed selection. Sometimes, it may also be possible to increase the loading or unloading rate with a proper incentive.

The problem of whether to accept a single cargo request or not is also an operational problem since the potential customer often requires an answer immediately, see Section 5.5. In practice, a cargo is often accepted if there is available capacity. However, accepting a new cargo may restrict the possibilities for taking a more profitable cargo that becomes available in the market later. Therefore, it could be advantageous to introduce the concept of stochastic optimization to such problems. The authors are not aware of such contributions.

6.4 Persistence

Schedules have often to be changed due to unforeseen delays, changes in requirements or other events. In such circumstances it may be highly desirable to minimize changes to already published schedules. Thus, necessary changes in the schedule of one vessel should have a minimal effect on the schedule of other vessels. Optimization models have a well-deserved reputation for amplifying small input changes into drastically different solutions. A previously optimal solution may still be nearly optimal in a new scenario and managerially preferable to a dramatically different solution that is mathematically optimal. Optimization models can be stated so that they exhibit varying degrees of persistence with respect to previous values of variables, constraints, or even exogenous considerations. Brown et al. (1997a) discuss these aspects of optimization and persistence.

In another paper by Brown et al. (1997b), the persistence aspect is considered when optimizing submarine berthing plans. Once in port, submarines may

be shifted to different berthing locations to allow them to better receive services that they require, or to clear space for other shifted vessels. Submarine berth shifting is expensive, labor intensive and may be potentially hazardous. Brown et al. (1997b) present a mixed-integer programming model for this berth planning problem with a planning horizon of 1–2 weeks. Once a berthing plan has been approved, changes are inevitable due to delays, changed requests for services, and early arrival of inbound submarines. An optimization model that only minimizes the costly berth shifts is not appropriate in this situation, because it can amplify minor modifications in service requests into wholesale revisions in the approved berthing plan. Revisions to the plan and the disruptions they bring must therefore be controlled to encourage trust in the planning process. Therefore Brown et al. (1997b) have incorporated a persistence incentive into the mixed-integer programming model that results in a decreased number of changes in previously published plans.

7 Perspectives and future research

As mentioned in the Introduction, demand for maritime transport services is increasing consistently, and there are no signs that the world economy will rely less heavily on maritime transport in the future. In this section we shortly discuss some trends in ocean shipping that will probably influence both the need for optimization-based decision support systems for maritime applications, and the shipping industry's acceptance of and benefits from such systems. We also wish to point out trends that result in a need for researchers to pay attention to new problem areas in maritime transportation. The focus is on applications within ship routing and scheduling. Trends in the land-side of maritime transportation operations are discussed in the Perspectives section of Crainic and Kim (2007). There may be additional trends, but these are the ones that we deem to be the primary ones, and that may have significant impact on the various aspects discussed in this chapter. A more detailed discussion of current trends in ship routing and scheduling is provided in Christiansen et al. (2004).

7.1 Mergers, acquisitions, and collaborations

During the last couple of decades we have witnessed consolidation in the manufacturing sector resulting in bigger actors on the demand side for maritime transport services. This has given the shippers increased market power compared to the shipping companies, resulting in squeezed profit margins for the shipping companies. In order to reduce this imbalance, there have been many mergers among shipping companies in the last decade. Many shipping companies have entered into pooling and collaboration efforts in order to increase their market power and gain flexibility in the services that can be offered (see Sheppard and Seidman, 2001). In such collaboration, a number of shipping companies bring their fleets into a pool and operate them together. The

income and costs are distributed among the different shipping companies according to certain rules that have been agreed upon. The split of income and costs is an intriguing topic for research.

Traditionally, scheduling in maritime transportation has been done manually by pencil and paper, based on the planners' knowledge and experience. The above trends of mergers and pooling collaborations result in larger controlled fleets. This means that it becomes much harder to determine a fleet schedule only by manual planning methods. Therefore, the need for optimization-based decision support systems has increased and will probably continue to increase in the future.

7.2 New generation of planners

Decision-makers and planners in the shipping industry are traditionally experienced, often with a sea-going background. As the fleets become larger, the planning problems focused on in this chapter become much harder to handle by manual methods. Despite this, planners are often very skeptical of computers in general and of optimization-based decision support systems in particular. However, in recent years we have seen that shipping companies have started employing planners with less practical but more academic background. This new generation of planners is more used to computers and software, and therefore is often much more open to new ideas such as using optimization-based decision support systems for the different applications in maritime transportation. Even though there is still a gap to bridge between researchers and planners in the shipping industry, we expect more willingness and interest from the ocean shipping industry to introduce such systems in the future.

7.3 Developments in software and hardware

The fast technological development in computers and communications also weighs heavily for the introduction of optimization-based decision support systems in shipping companies. Many earlier attempts failed due to restricted computer power, making it hard to model all the important problem characteristics and to facilitate a good user interface. However, today's computers enable an intuitive user interface to be implemented, something that is crucial for acceptance by the planners. In addition, there have been significant algorithmic developments. This, together with advances in computing power, has made it feasible to find good solutions to hard problems in a reasonable amount of time.

7.4 Shift from industrial to tramp shipping

Looking at the literature review on ship routing and scheduling presented by Christiansen et al. (2004), we observe that most contributions are in industrial shipping, while only a few are in the tramp market. In industrial shipping

the shipper controls the cargo and the fleet of ships. The purpose of an industrial operation is usually to provide the required transportation services for the organization's cargo requests at minimum cost. Industrial shipping is practiced by large extracting and manufacturing companies that have their own division that controls a number of ships for the transportation of their own cargoes. However, in recent years this has changed. Many such companies are now focusing on their core business and have outsourced other activities like transportation to independent shipping companies. Therefore, the emphasis has shifted somewhat from industrial to tramp shipping. Increasing global competition results in shifting industrial shipping operations from being considered "cost centers" into "profit centers" and compels them to become more involved in the spot market. This also brings new opportunities for optimization-based decision support systems for ship scheduling planners.

7.5 Focus on supply chains

In most ship scheduling studies reported in the literature, the supply chain perspective is missing. Recently we see an increasing competition between supply chains even more than between shipping companies. Shipping companies must consider themselves as total logistics providers, or at least as a part of a total logistics provider, instead of only a provider of sea transport services. This means that there must be some sort of collaboration and integration along the supply chain, for instance, between the shippers and the shipping company. Vendor managed inventory takes advantage of the benefits of introducing this integration and transfers inventory management and ordering responsibilities completely to the vendor or the logistics provider. The logistics provider determines both the quantity and timing of customer deliveries. The customer is guaranteed not to run out of product, and in return the logistics provider gains a dramatic increase in flexibility that leads to more efficient use of its resources.

We expect an increasing emphasis on integrating maritime transportation into the supply chain. This will also bring new interesting challenges to the research community in routing and scheduling, such as inventory routing, collaboration, and cost and/or profit sharing along the supply chain.

7.6 Strategic planning issues and market interaction

Vessel fleet sizing should be given more attention in the future. This strategic problem is extremely important as decisions concerning fleet size and composition set the stage for routing and scheduling. Even though there have been a few studies on this type of problem, the potential for improving fleet size decisions by using optimization-based decision support systems is probably significant. As already discussed, we have seen a trend from industrial to tramp shipping, with much more interaction with the market. This high degree of market interaction probably makes the fleet size issue even more important

and complex, as one now has to make some assumptions on future market development in order to determine the optimal fleet.

Contract evaluation (discussed in Section 3.5) is yet another important strategic problem that has only scarcely been considered in the research literature. This is to a large extent related to the fleet size issue, since the shipping company has to evaluate whether it has sufficient fleet tonnage to fulfill potential contract commitments together with its existing commitments. If so, one has to check whether a contract is profitable or not. In order to do so, one also has to make some assumptions about how the spot market will develop for the given contract period. Since both fleet sizing and contract evaluation decisions are to a large extent dependent on the expectations of how a future market will develop, concepts of optimization under uncertainty must probably be considered.

8 Conclusion

Maritime transportation is the backbone of international trade. The volume of maritime transportation has been growing for many years, and is expected to continue growing in the foreseeable future. Maritime transportation is a unique transportation mode possessing characteristics that differ from other modes of transportation, and requires decision support models that fit the specific problem characteristics.

Maritime transportation poses a rich spectrum of decision making problems, from strategic ones through tactical to operational. We also find within maritime transportation a variety of modes and types of operations with their specific characteristics: industrial, tramp, liner, deep-sea, short-sea, coastal and inland waterways, port and container terminals, and their interface with vessels.

Research interest in maritime transportation problems has been increasing in recent years but still lags behind the more visible modes, namely truck, air, and rail. In this chapter we have presented a variety of decision making problems in maritime transportation. For some common problems we presented models as well as discussed solution approaches, whereas for other problems we confined ourselves to a general description of the problems and referred the reader to sources that deal with the problems more extensively. Although most of the research in maritime transportation stemmed from real-life problems only a fraction of it has matured into real decision support systems that are used in practice.

The fast containerization of general and break-bulk cargo combined with fast development of information technology and telecommunications, and with competitive pressures, have resulted in a shift of emphasis from ocean transportation to intermodal supply chains. The economies of scale that such supply chains pose result in industry consolidation and larger controlled fleets, presenting a fertile ground for applying quantitative decision support tools. At the

same time shippers started to focus on their core operations and to outsource logistic functions to third party providers who also have significant economies of scale. Thus, also on the demand side we observe consolidation and higher potential for applying quantitative decision support tools.

Uncertainty plays a major role in maritime transportation and therefore robust and stochastic models should take center stage. However, in this respect the surface has only been scratched.

Maritime transportation poses a wide variety of challenging research problems, the solutions to which have high potential to improve economic performance and increase profitability in this highly competitive arena. The fast development of optimization algorithms and computing power facilitate solution of more realistic problems, and we are confident that more research will be directed to this crucial transportation mode.

Acknowledgements

This work was carried out with financial support from the Research Council of Norway through the TOP project (Improved Optimisation Methods in Transportation Logistics), the INSUMAR project (Integrated supply chain and maritime transportation planning) and the OPTIMAR project (Optimization in Maritime transportation and logistics). We want to thank the Doctoral students Roar Grønhaug, Frank Hennig, and Yuriy Maxymovych for a careful reading of the chapter and for helpful suggestions.

References

Almogy, Y., Levin, O. (1970). Parametric analysis of a multi-stage stochastic shipping problem. In: *IFORS, OR 69: Proc. Fifth International Conference on OR*. Tavistock, London, pp. 359–370.

Appelgren, L.H. (1969). A column generation algorithm for a ship scheduling problem. *Transportation Science* 3, 53–68.

Appelgren, L.H. (1971). Integer programming methods for a vessel scheduling problem. *Transportation Science* 5, 64–78.

Ariel, A. (1991). The effect of inventory holding costs on the optimal payload of bulk carriers. *Maritime Policy & Management* 18 (3), 217–224.

Avriel, M., Penn, M., Shipier, N., Witteboon, S. (1998). Stowage planning for container ships to reduce the number of shifts. *Annals of Operations Research* 76, 55–71.

Bausch, D.O., Brown, G.G., Ronen, D. (1998). Scheduling short-term marine transport of bulk products. *Maritime Policy & Management* 25 (4), 335–348.

Bellmore, M. (1968). A maximum utility solution to a vehicle constrained tanker scheduling problem. *Naval Research Logistics Quarterly* 15, 403–411.

Bendall, H.B., Stent, A.F. (2001). A scheduling model for a high speed containership service: A hub and spoke short-sea application. *International Journal of Maritime Economics* 3 (3), 262–277.

Brown, G.G., Graves, G.W., Ronen, D. (1987). Scheduling ocean transportation of crude oil. *Management Science* 33 (3), 335–346.

Brown, G.G., Goodman, C.E., Wood, R.K. (1990). Annual scheduling of Atlantic Fleet naval combatants. *Operations Research* 38 (2), 249–259.

Brown, G.G., Dell, R.F., Farmer, R.A. (1996). Scheduling coast guard district cutters. *Interfaces* 26 (2), 59–72.

Brown, G.G., Dell, R.F., Wood, R.K. (1997a). Optimization and persistence. *Interfaces* 27 (5), 15–37.

Brown, G.G., Cormican, K.J., Lawphongpanich, S., Widdis, D.B. (1997b). Optimizing submarine berthing with a persistence incentive. *Naval Research Logistics* 44, 301–318.

Brønmo, G., Christiansen, M., Nygreen, B. (2006). Ship routing and scheduling with flexible cargo sizes. *Journal of the Operation Research Society*, doi:10.1057/palgrave.jors.2602263. Advance online publication, 16 August 2006.

Brønmo, G., Christiansen, M., Fagerholt, K., Nygreen, B. (2007). A multi-start local search heuristic for ship scheduling – a computational study. *Computers & Operations Research* 34 (3), 900–917.

Chajakis, E.D. (1997). Sophisticated crude transportation. *OR/MS Today* 24 (6), 30–34.

Chajakis, E.D. (2000). Management science for marine petroleum logistics. In: Zanakis, S.H., Doukidis, G., Zopounidis, C. (Eds.), *Decision Making: Recent Developments and Worldwide Applications*. Kluwer Academic, pp. 169–185.

Cho, S.-C., Perakis, A.N. (1996). Optimal liner fleet routing strategies. *Maritime Policy & Management* 23 (3), 249–259.

Cho, S.-C., Perakis, A.N. (2001). An improved formulation for bulk cargo ship scheduling with a single loading port. *Maritime Policy & Management* 28 (4), 339–345.

Christiansen, M. (1999). Decomposition of a combined inventory and time constrained ship routing problem. *Transportation Science* 33 (1), 3–16.

Christiansen, M., Fagerholt, K. (2002). Robust ship scheduling with multiple time windows. *Naval Research Logistics* 49 (6), 611–625.

Christiansen, M., Nygreen, B. (1998a). A method for solving ship routing problems with inventory constraints. *Annals of Operations Research* 81, 357–378.

Christiansen, M., Nygreen, B. (1998b). Modeling path flows for a combined ship routing and inventory management problem. *Annals of Operations Research* 82, 391–412.

Christiansen, M., Nygreen, B. (2005). Robust inventory ship routing by column generation. In: Desaulniers, G., Desrosiers, J., Solomon, M.M. (Eds.), *Column Generation*. Springer-Verlag, New York, pp. 197–224.

Christiansen, M., Fagerholt, K., Ronen, D. (2004). Ship routing and scheduling: Status and perspectives. *Transportation Science* 38 (1), 1–18.

Cline, A.K., King, D.H., Meyering, J.M. (1992). Routing and scheduling of coast guard buoy tenders. *Interfaces* 22 (3), 56–72.

Crainic, T.G., Kim, K.H. (2007). Intermodal transportation. In: Barnhart, C., Laporte, G. (Eds.), *Transportation*. *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, pp. 467–537. This volume.

Crary, M., Nozick, L.K., Whitaker, L.R. (2002). Sizing the US destroyer fleet. *European Journal of Operational Research* 136, 680–695.

Cullinane, K., Khanna, M. (1999). Economies of scale in large container ships. *Journal of Transport Economics and Policy* 33 (2), 185–208.

Dantzig, G.B., Fulkerson, D.R. (1954). Minimizing the number of tankers to meet a fixed schedule. *Naval Research Logistics Quarterly* 1, 217–222.

Darby-Dowman, K., Fink, R.K., Mitra, G., Smith, J.W. (1995). An intelligent system for US coast guard cutter scheduling. *European Journal of Operational Research* 87, 574–585.

Darzentas, J., Spyrou, T. (1996). Ferry traffic in the Aegean Islands: A simulation study. *Journal of the Operational Research Society* 47, 203–216.

Desrosiers, J., Dumas, Y., Solomon, M.M., Soumis, F. (1995). Time constrained routing and scheduling. In: Ball, M.O., Magnanti, T.L., Monma, C.L., Nemhauser, G.L. (Eds.), *Network Routing*. *Handbooks in Operations Research and Management Science*, vol. 8. North-Holland, Amsterdam, pp. 35–139.

Erkut, E., Verter, V. (2007). Hazardous materials transportation. In: Barnhart, C., Laporte, G. (Eds.), *Transportation*. *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, pp. 539–621. This volume.

European Commission (2004). European Transport Policy for 2010: Time to Decide. White paper.

Fagerholt, K. (1999). Optimal fleet design in a ship routing problem. *International Transactions in Operational Research* 6 (5), 453–464.
Fagerholt, K. (2001). Ship scheduling with soft time windows – an optimization based approach. *European Journal of Operational Research* 131, 559–571.
Fagerholt, K. (2004). A computer-based decision support system for vessel fleet scheduling – experience and future research. *Decision Support Systems* 37 (1), 35–47.
Fagerholt, K., Christiansen, M. (2000a). A combined ship scheduling and allocation problem. *Journal of the Operational Research Society* 51 (7), 834–842.
Fagerholt, K., Christiansen, M. (2000b). A travelling salesman problem with allocation, time window and precedence constraints – an application to ship scheduling. *International Transactions in Operational Research* 7 (3), 231–244.
Fagerholt, K., Lindstad, H. (2000). Optimal policies for maintaining a supply service in the Norwegian Sea. *OMEGA* 28, 269–275.
Fagerholt, K., Rygh, B. (2002). Design of a sea-borne system for fresh water transport – A simulation study. *Belgian Journal of Operations Research, Statistics and Computer Science* 40 (3–4), 137–146.
Fisher, M.L., Rosenwein, M.B. (1989). An interactive optimization system for bulk-cargo ship scheduling. *Naval Research Logistics* 36, 27–42.
Flatberg, T., Haavardtun, H., Kloster, O., Løkketangen, A. (2000). Combining exact and heuristic methods for solving a vessel routing problem with inventory constraints and time windows. *Ricerca Operativa* 29 (91), 55–68.
Fleming, D.K. (2002). Reflections on the history of US cargo liner service (part I). *International Journal of Maritime Economics* 4 (4), 369–389.
Fleming, D.K. (2003). Reflections on the history of US cargo liner service (part II). *Maritime Economics & Logistics* 5 (1), 70–89.
Fox, M., Herden, D. (1999). Ship scheduling of fertilizer products. *OR Insight* 12 (2), 21–28.
Garrod, P., Miklius, M. (1985). The optimal ship size: A comment. *Journal of Transport Economics and Policy* 19 (1), 83–91.
Gillman, S. (1999). The size economies and network efficiencies of large containerships. *International Journal of Maritime Economics* 1 (1), 39–59.
Hersh, M., Ladany, S.P. (1989). Optimal scheduling of ocean cruises. *INFOR* 27 (1), 48–57.
Hughes, W.P. (2002). Navy operations research. *Operations Research* 50 (1), 103–111.
Hwang, S.-J. (2005). Inventory constrained maritime routing and scheduling for multi-commodity liquid bulk. Phd thesis, Georgia Institute of technology, Atlanta.
Jaikumar, R., Solomon, M.M. (1987). The tug fleet size problem for barge line operations: A polynomial algorithm. *Transportation Science* 21 (4), 264–272.
Jansson, J.O., Shneerson, D. (1978). Economies of scale of general cargo ships. *The Review of Economics and Statistics* 60 (2), 287–293.
Jansson, J.O., Shneerson, D. (1982). The optimal ship size. *Journal of Transport Economics and Policy* 16 (3), 217–238.
Jansson, J.O., Shneerson, D. (1985). A model of scheduled liner freight services: Balancing inventory cost against shipowners' costs. *The Logistics and Transportation Review* 21 (3), 195–215.
Jansson, J.O., Shneerson, D. (1987). *Liner Shipping Economics*. Chapman & Hall, London.
Jaramillo, D.I., Perakis, A.N. (1991). Fleet deployment optimization for liner shipping. Part 2. Implementation and results. *Maritime Policy & Management* 18 (4), 235–262.
Jetlund, A.S., Karimi, I.A. (2004). Improving the logistics of multi-compartment chemical tankers. *Computers & Chemical Engineering* 28, 1267–1283.
Kang, J.-G., Kim, Y.-D. (2002). Stowage planning in maritime container transportation. *Journal of the Operational Research Society* 53, 415–426.
Kao, C., Chen, C.Y., Lyu, J. (1993). Determination of optimal shipping policy by inventory theory. *International Journal of Systems Science* 24 (7), 1265–1273.
Kleywegt, A. (2003). Contract planning models for ocean carriers. Working paper, Georgia Institute of Technology, Atlanta, GA.
Korsvik, J.E., Fagerholt, K., Brønmo, G. (2007). Ship scheduling with flexible cargo quantities: A heuristic solution approach. Working paper, Norwegian University of Science and Technology, Trondheim, Norway.

Ladany, S.P., Arbel, A. (1991). Optimal cruise-liner passenger cabin pricing policy. *European Journal of Operational Research* 55, 136–147.
Lane, D.E., Heaver, T.D., Uyeno, D. (1987). Planning and scheduling for efficiency in liner shipping. *Maritime Policy & Management* 14 (2), 109–125.
Larson, R.C. (1988). Transporting sludge to the 106-Mile site: An inventory/routing model for fleet sizing and logistics system design. *Transportation Science* 22 (3), 186–198.
Lawrence, S.A. (1972). *International Sea Transport: The Years Ahead*. Lexington Books, Lexington, MA.
Liu, C.-M., Sherali, H.D. (2000). A coal shipping and blending problem for an electric utility company. *OMEGA* 28, 433–444.
Lo, H.K., McCord, M.R., Wall, C.K. (1991). Value of ocean current information for strategic routing. *European Journal of Operational Research* 55, 124–135.
Martin, G.L., Randhawa, S.U., McDowell, E.D. (1988). Computerized container-ship loading: A methodology and evaluation. *Computers & Industrial Engineering* 14 (4), 429–440.
McCord, M.R., Lee, Y.-K., Lo, H.K. (1999). Ship routing through altimetry-derived ocean currents. *Transportation Science* 33 (1), 49–67.
McLellan, R.G. (1997). Bigger vessels: How big is too big? *Maritime Policy & Management* 24 (2), 193–211.
Mehrez, A., Hung, M.S., Ahn, B.H. (1995). An industrial ocean-cargo shipping problem. *Decision Sciences* 26 (3), 395–423.
Nulty, W.G., Ratliff, H.D. (1991). Interactive optimization methodology for fleet scheduling. *Naval Research Logistics* 38, 669–677.
O'Brien, G.G., Crane, R.R. (1959). The scheduling of a barge line. *Operations Research* 7, 561–570.
Papadakis, N.A., Perakis, A.N. (1989). A nonlinear approach to multiorigin, multidestination fleet deployment problem. *Naval Research Logistics* 36, 515–528.
Papadakis, N.A., Perakis, A.N. (1990). Deterministic minimal time vessel routing. *Operations Research* 38 (3), 426–438.
Perakis, A.N. (1985). A second look at fleet deployment. *Maritime Policy & Management* 12, 209–214.
Perakis, A.N., Bremer, W.M. (1992). An operational tanker scheduling optimization system. Background, current practice and model formulation. *Maritime Policy & Management* 19 (3), 177–187.
Perakis, A.N., Inozu, B. (1991). Optimal maintenance, repair, and replacement for Great Lakes marine diesels. *European Journal of Operational Research* 55, 165–182.
Perakis, A.N., Jaramillo, D.I. (1991). Fleet deployment optimization for liner shipping. Part 1. Background, problem formulation and solution approaches. *Maritime Policy & Management* 18 (3), 183–200.
Perakis, A.N., Papadakis, N.A. (1987a). Fleet deployment optimization models. Part 1. *Maritime Policy & Management* 14, 127–144.
Perakis, A.N., Papadakis, N.A. (1987b). Fleet deployment optimization models. Part 2. *Maritime Policy & Management* 14, 145–155.
Perakis, A.N., Papadakis, N.A. (1989). Minimal time vessel routing in a time-dependent environment. *Transportation Science* 23, 266–276.
Persson, J.A., Göthe-Lundgren, M. (2005). Shipment planning at oil refineries using column generation and valid inequalities. *European Journal of Operational Research* 163, 631–652.
Pesenti, R. (1995). Hierarchical resource planning for shipping companies. *European Journal of Operational Research* 86, 91–102.
Pope, J.A., Talley, W.K. (1988). Inventory costs and optimal ship size. *Logistics and Transportation Review* 24 (2), 107–120.
Powell, B.J., Perakis, A.N. (1997). Fleet deployment optimization for liner shipping: An integer programming model. *Maritime Policy & Management* 24 (2), 183–192.
Psaraftis, H.N. (1988). Dynamic vehicle routing problems. In: Golden, B.L., Assad, A.A. (Eds.), *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, pp. 223–248.
Psaraftis, H.N. (1999). Foreword to the focused issue on maritime transportation. *Transportation Science* 33 (1), 1–2.
Rana, K., Vickson, R.G. (1988). A model and solution algorithm for optimal routing of a time-chartered containership. *Transportation Science* 22 (2), 83–95.

Rana, K., Vickson, R.G. (1991). Routing container ships using Lagrangean relaxation and decomposition. *Transportation Science* 25 (3), 201–214.
Richetta, O., Larson, R. (1997). Modeling the increased complexity of New York City's refuse marine transport system. *Transportation Science* 31 (3), 272–293.
Ronen, D. (1982). The effect of oil price on the optimal speed of ships. *Journal of the Operational Research Society* 33, 1035–1040.
Ronen, D. (1983). Cargo ships routing and scheduling: Survey of models and problems. *European Journal of Operational Research* 12, 119–126.
Ronen, D. (1986). Short-term scheduling of vessels for shipping bulk or semi-bulk commodities originating in a single area. *Operations Research* 34 (1), 164–173.
Ronen, D. (1991). Editorial to the feature issue on water transportation. *European Journal of Operational Research* 55 (2), 123.
Ronen, D. (1993). Ship scheduling: The last decade. *European Journal of Operational Research* 71, 325–333.
Ronen, D. (2002). Marine inventory routing: Shipments planning. *Journal of the Operational Research Society* 53, 108–114.
Sambracos, E., Paravantis, J.A., Tarantilis, C.D., Kiranoudis, C.T. (2004). Dispatching of small containers via coastal freight liners: The case of the Aegean sea. *European Journal of Operational Research* 152, 365–381.
Schrady, D., Wadsworth, D. (1991). Naval combat logistics support system. *Journal of the Operational Research Society* 42 (11), 941–948.
Schwartz, N.L. (1968). Discrete programs for moving known cargos from origins to destinations on time at minimum bargeline fleet cost. *Transportation Science* 2, 134–145.
Scott, J.L. (1995). A transportation model, its development and application to a ship scheduling problem. *Asia-Pacific Journal of Operational Research* 12, 111–128.
Sheppard, E.J., Seidman, D. (2001). Ocean shipping alliances: The wave of the future? *International Journal of Maritime Economics* 3, 351–367.
Sherali, H.D., Al-Yakoob, S.M., Hassan, M.M. (1999). Fleet management models and algorithms for an oil tanker routing and scheduling problem. *IIE Transactions* 31, 395–406.
Shih, L.-H. (1997). Planning of fuel coal imports using a mixed integer programming method. *International Journal of Production Economics* 51, 243–249.
Sigurd, M.M., Ulstein, N.L., Nygreen, B., Ryan, D.M. (2005). Ship scheduling with recurring visits and visit separation requirements. In: Desaulniers, G., Desrosiers, J., Solomon, M.M. (Eds.), *Column Generation*. Springer-Verlag, New York, pp. 225–245.
Talley, W.K., Agarwal, V.B., Breakfield, J.W. (1986). Economics of density of ocean tanker ships. *Journal of Transport Economics and Policy* 20 (1), 91–99.
Thompson, P.M., Psaraftis, H.N. (1993). Cyclic transfer algorithms for multi-vehicle routing and scheduling problems. *Operations Research* 41 (5), 935–946.
UNCTAD (2003). *Review of Maritime Transport, 2003*. United Nations, New York and Geneva.
UNCTAD (2004). *Review of Maritime Transport, 2004*. United Nations, New York and Geneva.
Vis, I.F.A., de Koster, R. (2003). Transshipment of containers at a container terminal: An overview. *European Journal of Operational Research* 147, 1–16.
Vukadinovic, K., Teodorovic, D. (1994). A fuzzy approach to the vessel dispatching problem. *European Journal of Operational Research* 76 (1), 155–164.
Vukadinovic, K., Teodorovic, D., Pavkovic, G. (1997). A neural network approach to the vessel dispatching problem. *European Journal of Operational Research* 102 (3), 473–487.
Wermus, M., Pope, J.A. (1994). Scheduling harbor pilots. *Interfaces* 24 (2), 44–52.
Williams, H.P. (1999). *Model Building in Mathematical Programming*, 4th edition. Wiley, West Sussex, pp. 160–165.
Williams, T.M. (1992). Heuristic scheduling of ship replenishment at sea. *Journal of the Operational Research Society* 43 (1), 11–18.
Wilson, I.D., Roach, P.A. (2000). Container stowage planning: a methodology for generating computerized solutions. *Journal of the Operational Research Society* 51 (11), 1248–1255.
Xinlian, X., Tangfei, W., Daisong, C. (2000). A dynamic model and algorithm for fleet planning. *Maritime Policy & Management* 27 (1), 53–63.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14005-0

Chapter 5

Dynamic Models for Freight Transportation

Warren B. Powell

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA
E-mail: powell@princeton.edu

Belgacem Bouzaïene-Ayari

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA
E-mail: belgacem@princeton.edu

Hugo P. Simão

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA
E-mail: hpsimao@princeton.edu

1 Introduction

Dynamic models arise in a vast array of transportation applications as a result of the need to capture the evolution of activities over time. But exactly what do we mean by a "dynamic model"? All dynamic models capture the timing of physical activities which introduces both modeling and algorithmic challenges and opportunities. In real problems, there are three classes of activities that can evolve over time: the physical movement of freight, equipment and people; the evolution of information; and the movement of financial resources as services are paid for.

In this chapter we focus on modeling the organization and flow of information and decisions, in the context of freight transportation problems that involve the management of people and equipment to serve the needs of customers. The timing of the flow of capital is becoming an increasingly important dimension of freight transportation, but as of this writing, there has been virtually no formal research on the topic. Modeling the timing of physical activities, by contrast, dates to the 1950s. These models introduce a range of modeling and algorithmic challenges that have been studied since the early years of operations research models.

Our decision to focus on modeling the evolution of information (or more broadly, the organization and flow of information and decisions) reflects the growing maturity of this class of models, and the importance of questions that require an explicit model of information processes. Often categorized under the phrase "stochastic, dynamic models", there are two issues that arise in this context:

- Improving the management of physical resources by using a more accurate representation of information processes.
- Designing and controlling information processes to maximize performance.

The first class of questions focuses on more realistic models to improve the design and control of physical systems. Examples of questions include:

- How do we best allocate rail cars to handle future requests?
- Which driver should move a load of freight given the as-yet unknown loads that will be available at the destination of the first load?
- How many dock workers should be assigned to work a shift given the uncertain amount of freight to be unloaded?
- How many locomotives should be held at a yard to handle potential equipment failures of locomotives that are expected to arrive inbound over the next 12 hours?

All of these are examples of questions concerning the allocation of physical resources in the presence of uncertainty in the future. Most often, the uncertainty arises around demands that are placed on the system, but there can be uncertainty in the availability of resources (how many drivers can the trucking company hire this week? will the locomotive break down?) and the performance of the network itself (primarily travel times). A common theme when designing and controlling systems under uncertainty is producing solutions that work well over as broad a range as possible of potential outcomes in the future. Such solutions are referred to as *robust*, although this term is most often used in the context of design problems.

The second class of questions arises in the design and control of information processes. Examples include:

- Should a company invest in wireless communication devices that allow them to determine the location and status of drivers and equipment?
- Should a company invest in a database that tracks the status of all of its drivers/crew?
- What is the cost of allowing customers to make requests at the last minute?

Questions such as these require a model that allows us to explicitly manipulate the information process. If we consider a scenario which reduces the information available when we are making a decision, then we have to be sure that we are producing the best decisions that we can even if we do not have the information.

2 Some illustrative applications

It helps to begin our presentation with a discussion of different applications in freight transportation. Our presentation of models focuses on the resources

being managed, and the information and decision processes that govern the evolution of our systems. Below we list a series of application areas, along with examples of resource classes, (exogenous) information classes, and decision classes. For each subcategory, we briefly list typical examples of each class. For transportation problems, every resource includes as attributes its location and the estimated time at which the resource will arrive at the destination (later, we provide a more formal and general characterization of "estimated time of arrival"). For this reason, we exclude these attributes in the discussion below.

(1) Truckload trucking:

Resource classes:
- Drivers, who are normally modeled as including the tractor. Drivers are characterized by attributes such as location, home domicile, time due at home, international experience, and sleeper vs. single driver.
- Trailers: loaded status (loaded or empty), cleanliness, repair status.
- Loads of freight: destination, service requirements, equipment needs, shipper classification.

Information classes:
- Customer requests for service.
- Drivers entering and leaving the system.
- Equipment status.
- Transit times.

Decision classes:
- Drivers: couple with trailer and/or load, move empty, move loaded, go on rest (at home or away).
- Trailers: move (empty or loaded), clean, repair.
- Loads: accept/reject initial request, spot pricing of a load, couple with trailer, move (with trailer), move using a subcontractor (a driver not managed by the company).

(2) Rail car distribution:

Resource classes:
- Cars: car type, loaded status, maintenance status, cleanliness, shipper pool, owner.
- Orders: destination, service requirements, commodity type and substitutability.

Information classes:
- Customer orders, including updates to customer orders.
- Empty cars becoming available from shippers or other railroads.
- Transit times.
- The destination of an order (revealed only after a car is filled).
- Acceptability of a car to the customer (is it clean? in good repair?).

Decision classes:
- Cars: move (loaded or empty), repair, clean, switch pools, move offline to another railroad, spot pricing of an order.

- Orders: accept/reject, subcontract.

(3) Locomotive management:

Resource classes:
- Locomotives: type, horsepower, high/low adhesion, fuel status, due-in-shop date, inbound train that the locomotive last pulled (determines the other locomotives the unit is connected to).
- Trains: train type (grain, coal, merchandise, stack, priority), destination, departure time, tonnage, horsepower required per ton being pulled.

Information classes:
- Locomotives: maintenance status, arrivals to system, departures from system.
- Trains: tonnage, additions to schedule, cancellations (annulments), delays.

Decision classes:
- Locomotives: assignments to trains, assignment to shop, repositioning (moving without a train).
- Trains: decisions to delay trains.

(4) Less-than-truckload trucking:

Resource classes:
- Shipments: destination, service requirement, size (weight and density).
- Trailers: type (typically 28′ or 48′), current load (shipments on the trailer), destination, departure time.
- Drivers: domicile, bid vs. nonbid, sleeper vs. single, days away from home, hours of service.
- Tractors: type, status.

Information classes:
- Shipments entering the system.
- Changes in driver status.
- Transfer of shipments at sort facilities.

Decision classes:
- Where to run trailers direct.
- What trailer to put a shipment on.
- Coupling of trailers into loads to be moved (typically involves pairing 28′ trailers).
- Assignment of drivers to loads.

(5) Intermodal container operations:

Resource classes:
- Containers: type, loaded status, cleanliness, repair status, ownership.
- Loads: destination, weight, service requirements, routing restrictions.

Information classes:
- Customer orders.
- Transit times.
- Movement of containers.

Decision classes:
- Customer orders: accept/reject, move via subcontractor, spot pricing of an order.
- Containers: assignment to order, movement (loaded or empty), assignment to ship or train departures, intra-port management, routing and scheduling by truck.

(6) Air cargo:

Resource classes:
- Aircraft: type, configuration, load status, maintenance schedule.
- Requirements (loads of freight): destination, weight, size characteristics, passengers, pickup and delivery windows.
- Pilots: training, home location, hours worked, work schedule.

Information classes:
- Requests to move freight and passengers.
- Equipment failures (and random changes in status).
- Weather delays.

Decision classes:
- Move aircraft from one location to another.
- Reconfigure aircraft (to handle passengers or different types of cargo).
- Repair, refuel.
- Outsource customer requests.

3 A resource model

At the heart of our problems is the challenge of modeling resources. Although the focus of this chapter is information, we are interested in information in the context of managing resources for freight applications.

We begin in Section 3.1 by addressing the simple question: what is a resource? Section 3.2 introduces a basic resource model which gives us a general notational framework for modeling resource classes and attributes. Section 3.3 introduces the concept of resource layers, which allows us to handle combinations such as driver/trailer/load and locomotive/train; these are composite resources that take on new behaviors. Section 3.4 notes the interaction between the design of the attribute space, and the flow of information and decisions. Section 3.5 shows how our attribute notation can be used to describe six classes of resource management problems. Finally, Section 3.6 closes with a discussion of three perspectives of the "state" of our system.

3.1 What is a resource?

Any freight transportation system can be thought of as a problem in the management of resources. When we put it this way, we tend to think of resources as people and equipment, and sometimes fixed facilities. When we formulate optimization models, resources tend to appear as right-hand side constraints, which suggests that resources are anything that constrains the system. From the perspective of models, we can divide all data into information classes, where "resources" are information classes that constrain the system. It is useful to distinguish between static resource classes (such as terminals) which constrain the system, but which may not themselves be decision variables, and active resource classes (people, equipment) which we are managing.

When we adopt this more formal definition of a resource, we find that "demands" (also known as customers, tasks, requirements, loads, moves, pickups/deliveries, schedules) are also a form of "resource". Often, the distinction between "managing resources" (objects over which we have control) to "serve customers" (objects over which we have no control) is an artificial one that does not stand up to scrutiny (and certainly not in the formalism of a mathematical model).

A simple example illustrates why it is useful to use a broader definition of resources. Consider the problem faced by a truckload carrier of moving a load from city i to city j. The driver at i needs to return to his home domicile in city d. For this reason, the carrier assigns the driver to the load but moves both to an intermediate relay r, where the driver and load separate. The driver is then assigned to another load that moves him closer to his home, while the load is assigned to a new driver who will presumably move it toward its final destination (in industries such as long-haul less-than-truckload trucking, a load can move through four or five such relays before reaching its destination). During this time, we are trying to get the driver to his home and the load to its destination. While the attributes of each are different, the modeling issues are the same.

A working definition of a resource is *an information class that constrains the system*. This broad definition includes not only equipment and people, but also fixed assets such as truck terminals and railroad track which limits our ability to move freight. An *active* resource class is *an endogenously controllable information class that constrains the system*. Drivers, trailers, locomotives, and planes are all examples of active resource classes. A load of freight also limits our system (it limits how many trucks we can move loaded). In most models, a load is an example of a *passive* resource class since it constrains the system, but we cannot manage it directly (we can move a trailer with a load in it, but not a load without a trailer). Exceptions arise with companies that can contract out movements. In this case, we can effectively move a load without using any of the company-owned drivers or equipment. In such a setting, the load becomes an active resource class.

3.2 A basic resource model

Our resource model, then, begins with a listing of the resource classes:

$\mathcal{C}^{\mathrm{R}}$ = set of resource classes.

We describe a resource using a vector of attributes

a = vector of attributes which describe the state of a resource,

$\mathcal{A}^c$ = space of possible attribute vectors of a resource in class $c \in \mathcal{C}^{\mathrm{R}}$.

Attributes can be categorical (home domicile, equipment type, maintenance status) or numerical (hours worked, days until next maintenance stop, fuel level). Attribute vectors can be divided between static attributes (for example, the type of equipment or domicile of a driver) and dynamic attributes (location, maintenance status, hours of service). Often, the set of static attributes is combined to identify a resource type or commodity, while the set of dynamic attributes is the state. *Dynamic* resources have at least one dynamic attribute; *static* resources are characterized by attributes that are completely static.

We describe the status of all the resources in our system using:

R_{ta}^c = the number of resources in class $c \in \mathcal{C}^{\mathrm{R}}$
with attribute a at time t,

$R_t^c = \left(R_{ta}^c\right)_{a \in \mathcal{A}^c, c \in \mathcal{C}^{\mathrm{R}}}$.

Recall that one of our resource classes is the customer demands. It is common to model customer demands as the only source of exogenous information (new requests being made on the system) but there are many problems where there are exogenous updates to the other resources in the system (drivers being hired, equipment breaking down, equipment arriving exogenously to the system). We capture all these forms of information using

$\widehat{R}_{ta}$ = change in R_{ta} between $t-1$ and t
due to exogenous information,

$\widehat{R}_t = (\widehat{R}_{ta})_{a \in \mathcal{A}}$.

3.3 Resource layering

It is often the case that resources from different classes will be *coupled* to form a *layered resource*. For example:

- When managing locomotives, a set of locomotives will be assigned to a train to move cars from one location to the next. When a locomotive is attached to a train, it can only be removed at a cost, so the behavior of a locomotive at a location depends on whether it is attached to a train or not. Furthermore, the ability to route a locomotive back to its

home shop is very much affected by the attributes of a train that it is attached to (a railroad might easily decide not to detach a locomotive from a train that it is pulling, even if the destination of the train takes the locomotive away from its maintenance base).

- A cargo aircraft loaded with freight that breaks down will have a different behavior from an empty aircraft.
- Decisions about a business jet, with a particular pilot moving a passenger, that has just broken down will depend on the attributes of the aircraft, pilot and passenger.

We have to define the ways that resources may be coupled, which produces a set of resource layers. A layer, in general, is a resource class comprised of the coupling of one or more resource classes. We might have two resource classes (pilots and aircraft) and two resource layers. The first layer might be just the pilots (P) while the second might be aircraft with pilots ($A|P$).

To clearly distinguish between the resource and the resource layer, we use the same letter as a resource surrounded by a pair of parentheses to denote that resource layer. We let

$$
\begin{aligned}
&(A) = A|P \text{ denotes the aircraft layer,}\\
&(P) = P \text{ denotes the pilot layer,}\\
&\mathcal{L} = \text{a } \textit{layering}, \text{ representing the set of resource layers}\\
&\quad = \big((A), (P)\big)\\
&\quad = \{A|P, P\},\\
&l = \text{the index for resource layer,} \quad l \in \mathcal{L}, \text{ i.e., } l = A|P \text{ or } l = P.
\end{aligned}
$$

For this example, we would refer to the pilot layer as a primitive layer, while the aircraft layer is a composite layer. The attributes of each layer are given by

$$
\begin{aligned}
&\mathcal{A}^{(P)} = \mathcal{A}^{P}, \quad \text{the attribute space of a pilot layer,}\\
&a^{(P)} = a^{P}, \quad \text{the attribute vector of a pilot layer,}\\
&\mathcal{A}^{(A)} = \mathcal{A}^{A|P}\\
&\qquad = \mathcal{A}^{A} \times \mathcal{A}^{P}, \quad \text{the attribute space of an aircraft layer,}\\
&a^{(A)} = a^{A|P}\\
&\qquad = a^{A}|a^{P} \in \mathcal{A}^{(A)}, \quad \text{the attribute vector of an aircraft layer.}
\end{aligned}
$$

It is very common in transportation and logistics to model one or two-layer problems. A good example of a one-layer problem is the assignment of aircraft to a fixed schedule of flights (where the departure time of the flight is fixed). A two-layer problem would arise when assigning truck drivers to loads of freight, where we have to determine when the load should be moved (some loads have 24 hour windows or longer). This would be an example of a two-layer problem with a single active layer (the driver) and a single passive layer

(the load). In many companies, it is possible to outsource a load, which is equivalent to moving a load without a driver, giving us an instance of a two-layer problem with two active layers. A three-layer problem might arise when we have to simultaneously manage drivers, tractors, and trailers, or pilots, aircraft, and loads of cargo.

Published examples of three-layer problems are fairly rare, but a good example is the NRMO model for the military airlift problem (Morton et al., 1996; Baker et al., 2002) which tracks the flows of aircraft, crews and cargo in a military airlift problem.

3.4 Attributes, decisions, and information

The attribute space $\mathcal{A}$ should not reflect every possible state that a resource may occupy, but rather the states where the arrival of new information, and/or the ability to make a decision (which, after all, represents the arrival of new information) might affect the behavior of the resource. For example, consider a military aircraft which has to move from the United States to India through a series of three intermediate airbases (for refueling and potential repairs). We would need to represent the attributes of the aircraft at any point where new information might change the movement of the aircraft. For example, equipment breakdowns, congestion at the airbase or weather delays can affect the timing of the route. Furthermore, this new information might result in decisions being made to reroute the aircraft. In the simulation package AMOS, used by the air mobility command, it is common not to model new information or decisions while an aircraft is enroute, in which case we only would need to model the aircraft at the end of the route.

The role of information and decisions in the design of the attribute space is especially important in the presence of layered resources. It is clear the attribute space of a layered resource is much more complex than that of either primitive resource. The question is: what do we need to model? Consider a simplified model of a fleet management problem where a vehicle (truck, jet, container) is assigned to move a request from one location to another. There are no intermediate stops, and no decisions can be made about the equipment being managed until it has finished a task. We might refer to this as a "two-layer problem" but we would never need to actually model a layered resource. We do not make any decisions about a piece of equipment until it has finished a task. While there is a period of time when the equipment is coupled with the customer request, there is no need to explicitly model this.

3.5 Major problem classes

We can use our attribute vector to describe six major problem classes. These are:

- *Inventory problems* – $a = \{\cdot\}$. This basic problem arises in classical inventory theory, where there is a single type of product being held in inventory. The attribute a is a null vector.
- *Multiproduct inventory* – $a = \{k\}$, where $k \in \mathcal{K}$ is a set of resource types (product classes). The attribute a consists of a single, static element.
- *Single commodity flow problems* – $a = \{i\}$, where $i \in \mathcal{I}$ is a set of states or locations of a piece of equipment. An example is managing a fleet of identical trucks, where the only attribute of a truck is its current location.
- *Multicommodity flow* – $a = \{k, i\}$, where $k \in \mathcal{K}$ represents types of resources and $i \in \mathcal{I}$ is a set of locations or states.
- *Heterogeneous resource allocation problems* – $a = (a_1, \ldots, a_n)$. Here we have an n-dimensional attribute vector, where a_i is the ith attribute of a. These applications arise in the management of people and complex equipment.
- *Multilayered resource scheduling problems* – $a = \{a^{c_1}|a^{c_2}|\cdots|a^{c_n}\}$. Now the attribute vector is a concatenation of attribute vectors.

These six problem classes represent an increasing progression in the size of the attribute space. The problems with smaller attribute spaces are generally viewed as flow problems, even if we require integer solutions.

3.6 The states of our system

In dynamic models, it is common to talk of the "state of the system", but it is very hard to define a state formally. A review of virtually any of the major textbooks on dynamic programming will fail to produce a formal definition of the state of the system.

We find that there are three perspectives of state variables. The first is the attribute vector a which can be thought of as the state of a single resource. In fields such as crew scheduling which make use of column generation methods (Desrosiers et al., 1995; Vance et al., 1997; Desrochers and Soumis, 1989), the attribute vector is referred to as a label, and the determination of the best schedule for a crew is formulated as a dynamic program over the state space $\mathcal{A}$.

The second is the vector R_t which represents the resource state of the system (in most dynamic programming applications, R_t is viewed as the state variable). While the size of $\mathcal{A}$ can range from one to the millions (or larger), the number of possible values of R_t can be truly huge. Let $\mathcal{R}$ be the space of possible values of R_t (which we assume is discrete), and let $N = \sum_{a\in\mathcal{A}} R_{ta}$ be the total number of resources being managed. Then it is straightforward to show that

$$|\mathcal{R}| = \binom{N + |\mathcal{A}| - 1}{|\mathcal{A}| - 1}. \tag{1}$$

Even for small problems, this number can be extremely large (creating problems where it exceeds 10^{50} is not that difficult). For practical purposes, it is useful to refer to such spaces as "effectively infinite".

The third state variable might be referred to as the "information state", although a more precise definition is the "knowledge state". Consider a problem where there is a series of customer demands $\widehat{D}_t$ which vary from period to period. We might estimate the mean demand using a simple exponential smoothing model

$$\bar{\mu}_t = (1-\alpha)\bar{\mu}_{t-1} + \alpha\widehat{D}_t.$$

In addition, we might estimate the variance using

$$\bar{\sigma}_t^2 = (1-\alpha)\bar{\sigma}_{t-1}^2 + \alpha(\widehat{D}_t - \bar{\mu}_{t-1})^2.$$

At time t, $(\bar{\mu}_t, \bar{\sigma}_t^2)$ would be part of what we know at time t. Given $(\bar{\mu}_t, \bar{\sigma}_t^2)$, we have all the information we need from history about customer demands. Our state variable would then be represented by $S_t = (R_t, \bar{\mu}_t, \bar{\sigma}_t^2)$.

3.7 Bibliographic notes

The mathematical properties of pure networks were discovered in the 1950s (see Ford and Fulkerson, 1962, for a classic early description of networks), but it was in the 1970s that Glover, Klingman, and Kennington discovered and exploited the intersection of network algorithms and emerging computer science techniques to produce exceptionally fast (even in those days) algorithms for solving large networks (see Glover et al., 1974; Langley et al., 1974; Glover et al., 1992, summarizes this body of work). Network models required that all the flows represented a single equipment type (see, for example, White, 1972). Most problems in practice involve different types of equipment (such as box cars with different configurations) with some degree of substitution allowed to satisfy demands (a shipper might accept different car types). The numerous applications of these *multicommodity flow problems* produced a vast literature which sought to exploit the embedded network structure to overcome the otherwise large size of these problems (see, for example, Lasdon, 1970; Kennington, 1978; Kennington and Helgason, 1980; Assad, 1978). A significant development in the multicommodity literature was its adaptation to solving routing and scheduling problems for vehicles and crews, where the latter are typically characterized by complex vectors of attributes (Desrosiers et al., 1984; Lavoie et al., 1988; Desrochers and Soumis, 1989; Vance et al., 1997; Barnhart et al., 1998). These techniques involve solving a multiattribute dynamic program for a single crew; a master program then chooses the best from a set of potential schedules to produce an optimal overall schedule. The attribute vector appears only in the dynamic program solved in the subproblem; once the potential tours are generated, the master problem is typically a set partitioning or set covering problem which is a very special type of integer multicommodity flow problem. Powell et al. (2002) formally proposes the "heterogeneous resource allocation problem" which is a generalization of the classical multicommodity flow problem. The modeling framework in this chapter is based

on Powell et al. (2001) which introduces a general problem class termed the "dynamic resource transformation problem".

4 Modeling exogenous information processes

The modeling of the flow of information can be surprisingly subtle. In a deterministic model, we might try to solve a problem that looks like

$$\begin{aligned} &\min_x \sum_{t\in\mathcal{T}} c_t x_t \\ &\quad \text{subject to} \\ &\qquad A_t x_t - B_{t-1}x_{t-1} = \widehat{R}_t, \\ &\qquad x_t \geqslant 0. \end{aligned} \tag{2}$$

In a model such as this, we would understand that $x_t = (x_{tij})_{i,j\in\mathcal{I}}$, where x_{tij} is the flow from city i to city j starting at time t (and arriving at j after some travel time τ_{ij}). In other words, x_t is the vector of activities *happening* at time t. We would assume that c_t are the costs incurred in that time period. In other words, in a model such as this, the index t refers to when things are *happening*.

In a model that reflects dynamic information processes, we have to combine models of physical activities over time with models of information arriving over time. Consider a problem that arises in freight car distribution. If we were to look at an order in a historical file, we would simply see a movement from one location to another at a point in time. But if we were to capture the evolution of information that occurred before this physical event, we would see a series of information events as depicted in Figure 1, which reflects: (1) the initial call that there is an order requiring a car; (2) the amount of time required to move the empty car to the order; (3) the acceptance of the car by the shipper (who will sometimes reject a car because it is damaged or dirty); (4) the time required to fill the car, and (5) the destination of an order (which is not revealed until the car is loaded).

In our model above, the right-hand side constraint $\widehat{R}_t$ in Equation (2) would normally be interpreted as the new arrivals to the system at time t. However, we need to distinguish between the "phone call" representing the information about a new order and the physical arrival of a new order.

In the remainder of this section, we set up the fundamentals for modeling dynamic information processes. Section 4.1 begins by establishing our convention for modeling time. Then, Section 4.2 introduces a notational style for modeling information processes. Section 4.3 introduces the important concept of lagged information processes that arises in almost all transportation problems. Finally, Section 4.4 provides an introduction to an important mathematical concept used widely in the probability community for modeling information.

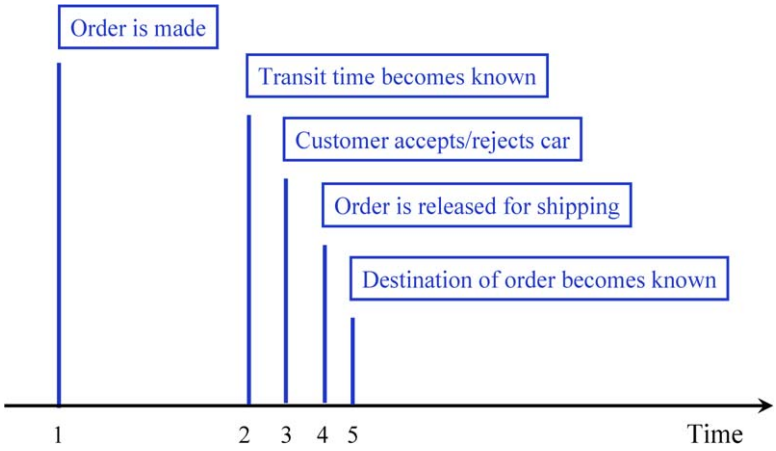


Fig. 1. The evolution of information for a single order in a freight car application.

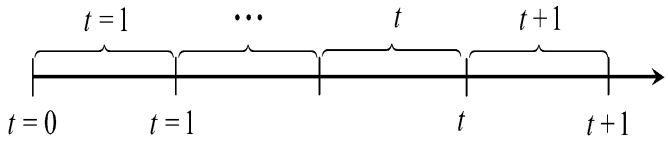


Fig. 2. The relationship between continuous and discrete time.

4.1 *Modeling time*

For computational reasons, it is common to model problems in transportation and logistics in discrete time, but it is apparent that in real problems, activities happen continuously. Surprisingly, the modeling of time (which is fairly obvious for deterministic problems) is not standard in the stochastic optimization community (for example, some authors like to index the first time period as $t = 0$ while others choose $t = 1$).

For computational reasons, we model our decision making process in discrete time (these can be fixed time periods such as hourly or daily, or driven by information events). However, it is useful to think of information arriving in continuous time, which resolves any ambiguity about what information is available when we make a decision. Figure 2 shows the relationship between discrete and continuous time which is most commonly used in the stochastic process community (Cinlar, 2003). In this convention, time $t = 0$ refers to "here and now", while discrete time period t refers to the time interval $(t-1, t]$.

It is very important to identify the information content of a variable. In a purely discrete time model, the precise sequencing of decisions and information is ambiguous (was a decision at time t made before or after the information that arrived at time t?). For this reason, we will view information as arriving continuously over time, while a decision will be made at a discrete

point in time. Since a state variable is the information we need to make a decision, it will also be measured in discrete time.

4.2 The information process

In the math programming community, it is common to use x_t as a generic decision variable (x_t might come in different flavors, or we might include y_t and z_t), but there is not a commonly accepted "generic variable" for information. For our presentation, we adopt the convention that W_t is a variable representing all the information that arrived during the time interval between $t-1$ and t. This information can be about new customer orders, transit times, costs, and the status of people and equipment. We may in general define

$$\begin{aligned}
\mathcal{C}^{\mathrm{I}} &= \text{the set of dynamic, exogenous information classes, representing} \\
&\qquad \text{different types of information arriving over time (prices,} \\
&\qquad \text{demands, equipment failures, travel times),} \\
W_t^c &= \text{the information arriving for class } c, \\
W_t &= \left(W_t^c\right)_{c\in\mathcal{C}^{\mathrm{I}}}.
\end{aligned}$$

In a particular application, it will be common to model specific information processes and give each of these names such as demands or prices. As a result, it may not be necessary to use the generic "W_t" notation. But it can be useful to have a generic information variable (as it will be in our presentation) since it will allow us to add new information processes without changing our basic model. We adopt the notation that with the exception of our generic information variable W_t, we use hats to indicate exogenous information processes. For example, if a trucking company faces random demands and spot prices, we might let $\widehat{D}_t$ and $\hat{p}_t$ be the demands and prices that we learn about during time interval t (between $t-1$ and t). Thus we would write $W_t = (\widehat{D}_t, \hat{p}_t)$. Given our convention for labeling time, W_1 is our first set of information arriving (W_0 has no meaning since information arrives over time intervals, and the first time interval is $t = 1$).

We further adopt the notation that any variable with subscript t uses information up through time t. This notation is of fundamental importance, and represents a significant departure from standard deterministic dynamic models where t represents when a physical activity is happening. Thus, x_t is a decision that uses information up through time t; S_t might be a state variable at time t (which uses W_t), and $C_t(x_t)$ is the cost computed using information up through time t.

Normally, models that capture dynamic information processes do so because we do not know what information is going to arrive in the future. This is fundamental to stochastic models, but we wish to emphasize that there are numerous engineering applications which explicitly model dynamic information, but model only one instance of information (for example, what might have occurred in history). When we wish to model multiple information processes, we

can adopt the standard notation of the probability community. Let ω represent a particular realization of the information process: $W_1, W_2, \ldots, W_t, \ldots$, and let Ω be the set of all possible realizations (we now see the motivation for using W_t as our generic variable for information – it "looks like" ω).

It is typically the case that we will be working with a single realization at a time. We can think of ω as an indexing of all the possible outcomes (as in the integers from 1 to $|\Omega|$), or as the literal outcomes of all the possible pieces of information that become known. When we are using Monte Carlo sampling techniques, we might think of ω as the random number seed to generate all the possible random variables. These are all effectively equivalent views, but we suggest that it is generally best to think of ω as a simple index. When we want to refer to the information that becomes known during time interval t, we may write $\omega_t = W_t(\omega)$. In this case, ω_t is a sample realization of all the different sources of exogenous information. If we view ω as the actual information in a realization, we can write

$$\omega = (\omega_1, \omega_2, \ldots, \omega_t, \ldots).$$

Some authors use ω or ω_t as if they are random variables, while the probability community insists that ω is strictly a sample realization. Readers new to the field need to understand that there is an astonishing lack of unanimity in notational styles. For example, the probability community will insist that an expression such as $Ef(x, \omega)$ has no meaning (ω is not random, so the expectation is not doing anything), while others have no objection to an expression such as this. The debate can be avoided by writing $Ef(x, W)$ where W is our random "information variable". Readers in engineering might feel that such debates are somewhat pedantic, but the issues become more relevant for researchers who wish to make fundamental theoretical contributions. Our position is to adopt the simplest possible notation that is as mathematically correct as possible.

In time-staged problems, it is often necessary to represent the information that we know up through time t. For this, we suggest the notation

$$\begin{aligned}
H_t &= \text{the history of the process, consisting of all the information} \\
&\quad \text{known through time } t \\
&= (W_1, W_2, \ldots, W_t), \\
\mathcal{H}_t &= \text{the set of all possible histories through time } t \\
&= \{H_t(\omega) \mid \omega \in \Omega\}, \\
h_t &= \text{a sample realization of a history} \\
&= H_t(\omega).
\end{aligned}$$

Later, we are going to need to define the subset of Ω that corresponds to a particular history. For this purpose, we define

$$\Omega_t(h_t) = \{\omega \mid (W_1(\omega), W_2(\omega), \ldots, W_t(\omega)) = h_t, \omega \in \Omega\}. \tag{3}$$

The notation for modeling the information up to a particular time period is not standard, and readers need to be prepared for a range of different notations in the literature on stochastic optimization.

Our interest is in systems where exogenous information and decisions occur sequentially. It might be useful for some readers to think of "decisions" as *endogenously controllable information classes*. If ω_t is a sample realization of the information arriving during time interval t, then x_t is the decision made at time t (using the information up through time t). Let S_t be the state of our system immediately before a decision is made (which means that it also contains ω_t). It is customary to represent the evolution of the state variable using

$$S_{t+1} = S^{\mathrm{M}}(S_t, x_t, \omega_{t+1}). \tag{4}$$

The superscript "M" refers to "model" since many authors refer to Equation (4) as the "plant model" or "system model". We note that some authors will write Equation (4) as $S_{t+1} = S^{\mathrm{M}}(S_t, x_t, \omega_t)$. This means that t refers to when the information is used, rather than the information content (we refer to this as an *actionable* representation, while Equation (4) is an *informational* representation). For simple problems, either representation works fine, but as richer problems are addressed, we believe that the informational representation will prove more effective.

The state variable S_t is the information available right before we make a decision, so we might call it the pre-decision state variable. For algorithmic purposes, we sometimes need to use the post-decision state variable (denoted S_t^x), which is the state right after we make a decision, before any new information arrives. The relationship between pre- and post-decision state variables is described using

$$S_t^x = S^{\mathrm{M},x}(S_t, x_t), \tag{5}$$

$$S_{t+1} = S^{\mathrm{M},W}\big(S_t^x, W_{t+1}\big). \tag{6}$$

This representation becomes particularly useful when using approximate dynamic programming methods (see Section 8.3). Finally, it is sometimes convenient to work only with the post-decision state variable. If this choice is made, we suggest simply defining S_t to be the post-decision state variable (without the superscript x), and writing the transition function using

$$S_t = S^{\mathrm{M}}(S_{t-1}, \omega_t, x_t).$$

Note that we are using the same function $S^{\mathrm{M}}(\cdot)$, but we change the order of the information and decision variables, reflecting the order in which they occur.

4.3 Lagged information processes

When we are solving deterministic models, we assume all information is known in advance. By contrast, most stochastic models assume that the information about an event (in particular a customer demand) becomes known at

the same time as the event actually happens. It is often the case, however, that the information about an event arrives at a different time than the physical event. The most common example is customers calling in orders in advance, but there are examples where the information arrives after an event (the dispatch of a truck might be entered into the computer hours after a truck has left).

We refer to these problems as instances of lagged information processes, reflecting the more common instance when information arrives before an event. This is quite common in freight transportation, and represents an important area of study. For example, carriers may be able to charge extra for last-minute requests, so they need to be able to estimate the cost of the higher level of uncertainty.

In general, a customer might call in a request for service, and then further modify (and possibly cancel) the request before it is finally served. For our purposes, we are going to treat the information in the initial phone call as accurate, so that when we receive information about a future event we can then model that event as being fully known. We introduce two terms that help us with these problems.

Knowable time – This is the time at which the information about a physical resource arrives.

Actionable time – This is the time at which a physical resource can be acted on.

We assume throughout that the knowable time arrives at or before the actionable time. The information process is assumed to bring to us the attributes of a resource that might not be actionable. This can be a customer booking an order in advance, or information that a vehicle has departed one location (the information event) headed for a destination in the future (at which point it becomes actionable).

The actionable time can be viewed as nothing more than an attribute of a resource. We can pick one element of the attribute vector a, and give it a name such as $a_{\text{actionable}}$. However, it is often useful to explicitly index the actionable time to simplify the process of writing flow conservation constraints, since it will usually be the case that we can only act on a resource (with a decision) that is actionable. For this reason, we let

$$\begin{aligned} R_{t,t'a} &= \text{the number of resources that are known at time } t \text{ to be} \\ &\qquad \text{actionable at time } t' \geqslant t \text{ with attribute vector } a, \\ R_{tt'} &= (R_{t,t'a})_{a\in\mathcal{A}}, \\ R_t &= (R_{tt'})_{t'\geqslant t}. \end{aligned}$$

Thus, we can continue to use R_t as the resource that we know about at time t. If we write R_{ta}, then we implicitly assume that the actionable time is an attribute. If we write $R_{tt'}$ or $R_{t,t'a}$ then we assume that t' is the actionable time.

The double time index (t, t') provides an explicit model of the information process (the first time index) and the physical process (the second time in-

dex). For deterministic models, we could (clumsily) index every variable by R_{0t}, whereas if we want to model processes where the information arrives as the physical event occurs, variables would all be indexed by R_{tt}. Clearly, in both cases it is more natural and compact to use a single time index.

The double time indexing can be used elsewhere. We might plan an activity at time t (implicitly, using the information available at time t) that will happen at time t' in the future. Such a decision vector would be denoted $x_{tt'}$, where x_{tt} is an action (something implemented right away) while $x_{tt'}$, $t' > t$, would represent a plan of an activity that might happen in the future. We can lock in such plans (normally we might do so for t' within some *decision horizon*), or treat them simply as forecasts of activities in the future (similar to forecasts of customer demands). As we did with R_t, we can let

$$x_t = (x_{tt'})_{t' \geqslant t},$$
$$c_t = (c_{tt'})_{t' \geqslant t}.$$

This allows us to retain a compact notation with a single time index, as long as it is understood that these variables can be vectors extending over a range of actionable times in the future.

Although we have not formally introduced our cost function, we can also let $c_{tt'}$ be the cost of an activity planned (using the information available) at time t. We note that the flow of dollars is itself a separate process from the flow of information and physical flows. Normally, payment for a service comes at some time after the service is provided (and as with the flow of the physical resource, there is some uncertainty associated with the financial flows). We are not going to further address the flow of money, but a complete dynamic model would separate the timing of the informational, physical and financial flows.

4.4 Information and sigma-algebras

Students of mathematical programming learn that they have to understand linear algebra and the meaning of equations such as $Ax = b$. People who work in the area of dynamic information processes (alternatively, stochastic, dynamic models) learn the meaning of statements such as "the function X_t must be $\mathcal{F}_t$-measurable". In this section, we undertake a brief explanation of this concept.

We first need to emphasize that measure theory is an abstract part of the field of mathematics that was adopted by the probability community (just as linear algebra was adopted by the math programming community). As a general rule, it is not necessary to have a strong command of measure theory to work successfully in the field of stochastic, dynamic models (in contrast with linear algebra which is essential to work in the field of math programming), but it will add richness to a presentation (and is essential for some theoretical work). Most important is the notion of a concept known as a "sigma-algebra" (often designated σ-algebra) and its role in conveying information. In engineering,

information is simply data; to a probabilist, information is always represented as a sigma-algebra.

If the random variables W_t were discrete, we could describe all the elements of Ω as "outcomes", with $p(\omega)$ the "probability of outcome ω". In general, this is not the case (random variables can be continuous), in which case we cannot refer to ω as an event. Instead, an event has to be a set of outcomes ω. For the moment (we refine this concept later), let $\mathcal{F}$ be all possible subsets of Ω, with the property that every union of sets in $\mathcal{F}$ is an element of $\mathcal{F}$, and every complement of $\mathcal{F}$ is a member of $\mathcal{F}$ (the set Ω and the empty set ϕ would both be elements of $\mathcal{F}$). Since every element of $\mathcal{F}$ is actually a set of events, $\mathcal{F}$ is, literally, a "set of sets". $\mathcal{F}$ is called a "sigma-algebra".

It is clear that for every set of outcomes Ω, we can define a sigma-algebra $\mathcal{F}$, and for every element in $\mathcal{F}$, we can define a probability. We let $\mathcal{P}$ be a function (which we will write as $p(\omega)$) defined over Ω that allows us to determine the probability of an event in $\mathcal{F}$ (if Ω is discrete, we just have to add up $p(\omega)$ for each ω in an event $\mathcal{E} \in \mathcal{F}$). If there are continuous random variables, then $p(\omega)$ will be a density function that we have to integrate over.

Given the set of outcomes Ω, events $\mathcal{F}$ and a probability measure, we have all the elements of a probability space which we write $(\Omega, \mathcal{F}, \mathcal{P})$. It is not uncommon to find papers which will start by defining a generic probability space $(\Omega, \mathcal{F}, \mathcal{P})$, without actually specifying what these are (nor do they refer to it again). For engineering applications, the outcomes Ω should be clearly defined in terms of actual information events.

This generic notation applies to any stochastic problem, but we are in particular interested in problems which involve information that arrives over time. Assume that we are at time t (the information during time interval $(t-1, t)$ has already arrived). We can define a sigma-algebra $\mathcal{F}_t$ consisting of all the events that can be created just using the information available up through time t. This is a fairly subtle concept that is best illustrated by an example. Assume we have a single customer which may or may not make a request in time period t. Our information W_t, then, is the random demand that we might call D_t (if this was our only source of exogenous information, we would just use the variable D_t for our information). In our simple example, D_t can be zero or one. If we are considering three time periods, then our sample space Ω consists of eight potential outcomes, shown in Table 1.

Now assume that we have just finished time interval 1. In this case, we only see the outcomes of D_1, which can be only 0 or 1. Let $\mathcal{E}_{\{D_1\}}$ be the set of outcomes ω that satisfy some logical condition on D_1. For example, $\mathcal{E}_{\{D_1=0\}} = \{\omega | D_1 = 0\} = \{1, 2, 3, 4\}$ is the set of outcomes that correspond to $D_1 = 0$. The sigma-algebra $\mathcal{F}_1$ would consist of the set of sets $\{\mathcal{E}_{\{D_1=0\}}, \mathcal{E}_{\{D_1=1\}}, \mathcal{E}_{\{D_1 \in (0,1)\}}, \mathcal{E}_{\{D_1 \notin (0,1)\}}\}$.

Next assume that we are in time period 2. Now our outcomes consist of the elementary events in $\mathcal{H}_2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, and the sigma-algebra $\mathcal{F}_2$ would include all the unions and complements of events in $\mathcal{H}_2$. For example, one event in $\mathcal{F}_2$ is $\{\omega | (D_1, D_2) = (0, 1)\} = \{3, 4\}$. Another event in $\mathcal{F}_2$ is

Table 1.
Set of demand outcomes

Outcome	Time period		
ω	1	2	3
1	0	0	0
2	0	0	1
3	0	1	0
4	0	1	1
5	1	0	0
6	1	0	1
7	1	1	0
8	1	1	1

the $\{\omega|(D_1, D_2) = (0, 0)\} = \{1, 2\}$. A third event in $\mathcal{F}_2$ is the union of these two events, which consists of $\omega = \{1, 2, 3, 4\}$ which, of course, is one of the events in $\mathcal{F}_1$. In fact, every event in $\mathcal{F}_1$ is an event in $\mathcal{F}_2$, but not the other way around. The reason is that the additional information from the second time period allows us to divide Ω into a finer set of subsets. Since $\mathcal{F}_2$ consists of all unions (and complements), we can always take the union of events which is the same as ignoring some piece of information. By contrast, we cannot divide $\mathcal{F}_1$ into a finer partition. The extra information in $\mathcal{F}_2$ allows us to filter Ω into a finer set of subsets than was possible when we only had the information through the first time period. If we are in time period 3, $\mathcal{F}$ will consist of each of the individual elements in Ω, as well as all the unions needed to create the same events in $\mathcal{F}_2$ and $\mathcal{F}_1$.

In other words, additional information allows us to divide our set of outcomes Ω into finer sets of subsets. This is the reason why a sigma-algebra is viewed as representing information.

With each additional time period, the information from that time period allows us to filter Ω into finer and finer subsets. As a result, we can write $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_t$. We term this a set of "increasing sub-sigma-algebras". When this occurs, we term the sequence $(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_t)$ a *filtration*. This is the reason why the notation $\mathcal{F}_t$ is used rather than a more mnemonic notation such as $\mathcal{H}_t$ (as in "history").

The sequence of sub-sigma-algebras is especially important, even if the mechanism will appear clumsy (not just at first – even experienced professionals feel this way). We can use this notation to specify that a sequential decision process is not cheating by using information in the future. When this is the case, we say that the decision rule is *nonanticipative*. Let X_t be a function that determines a decision at time t (more precisely, using the information available at time t). As further illustration, assume that we are faced with the problem of whether or not to dispatch a single truck with customers or freight waiting to move over a single link. In this case, $X_t \in (0, 1)$, where $X_t = 1$ means to dispatch the truck and $X_t = 0$ means to hold it another time period. We further

assume that the function X_t uses some well-defined rule to determine when the vehicle should be dispatched.

Our exogenous information process consists of A_t which is the number of customers arriving to move over our link. This problem is much too complex to use as a numerical example, but all the same concepts apply. However, this time we are going to define a set of events in a more meaningful way. Instead of a sigma-algebra that includes every possible event, we only care about the events that correspond to specific decisions. For example, for the decision X_1, we might assume that

$$X_1 = \begin{cases} 1, & \text{if } A_1 \geqslant y, \\ 0, & \text{otherwise,} \end{cases}$$

where y is a prespecified parameter. We do not care about every possible outcome of A_1; we only care about whether it is greater than or equal to y, or less than y. Let $\mathcal{E}_{\{A_1 \geqslant y\}} = \{\omega \mid A_1(\omega) \geqslant y\}$, with $\mathcal{E}_{\{A_1 < y\}}$ defined similarly. We can define a new sigma-algebra which we will call $\mathcal{F}_1^X$ which consists of the events $\{\mathcal{E}_{\{A_1 \geqslant y\}}, \mathcal{E}_{\{A_1 < y\}}, \Omega, \phi\}$. This is much smaller than $\mathcal{F}_1$, but contains the events that are relevant to our decision. We would say that $\mathcal{F}_1^X$ is the sigma-algebra *generated by* the decision function X_1. Clearly, $\mathcal{F}_1^X \subseteq \mathcal{F}_1$. Although it gets more complicated, we can continue this idea for other time periods. For example, we can think of a sequence of decisions (X_1, X_2, X_3). Each possible sequence, say $(0, 1, 0)$, corresponds to a set of realizations of (A_1, A_2, A_3). We can build a set of events $\mathcal{F}_3^X$ around all the possible sequences of decisions.

We have now built two sets of sigma-algebras. The original generic one, $\mathcal{F}_t$ which consists of all the possible events created by the sequence $(A_1, \ldots, A_t)$, and our new one, $\mathcal{F}_t^X$ which identifies the events relevant to our decision function, where $\mathcal{F}_t^X \subseteq \mathcal{F}_t$. Since we have a probability measure $\mathcal{P}$ defined on $(\Omega, \mathcal{F})$, we can compute the probability of any event in $\mathcal{F}_t$. This means that we can compute the probability of any event in $\mathcal{F}_t^X$. When this is the case, we say that "X_t is $\mathcal{F}_t$-measurable" because the sigma-algebra generated by the decision function X_t creates sets of events that are already elements of the original sigma-algebra (where we already know the probability of each event). It is common in stochastic optimization to create a decision function X_t and require it to be "$\mathcal{F}_t$-measurable". Equivalent terms are to say that X_t is $\mathcal{F}_t$-adapted, or that X_t is nonanticipative. All of these mean the same thing: that the decision function has access only to the information that has arrived by time t.

When would this not be the case? Let's say we cheated, and built X_t in a way that it used A_{t+1}. While this is not physically possible (you cannot use information that has not arrived yet), it is possible in computer simulations where, for example, we are solving a problem that happened in history. We might be running simulations to test new policies on historical data. It is possible in this situation to "cheat", which in formal terms could be called a "measurability violation". If a decision was able to see information in the future, we would be able to create a finer grained set $\mathcal{F}_t$. If some readers feel that this is a clumsy

way to communicate a basic idea, they would have a lot of company. But this is fairly standard vocabulary in the stochastic optimization community.

Another way to identify a valid set of decisions is to identify, for each outcome of the exogenous information ω (in our example, the realization of the number of arrivals in each time period), the corresponding set of decisions. Just as we can identify a set of events on the sequence $(A_1, A_2, \ldots, A_t)$, we can define sets of events on the sequence $(X_1, X_2, \ldots, X_t)$. For each $\omega \in \Omega$, there is a sequence of decisions $(x_1(\omega), x_2(\omega), \ldots, x_t(\omega))$. For lack of better notation, let $\mathcal{G}_t$ be the sigma-algebra created by all the possible outcomes of $(X_1, X_2, \ldots, X_t)$. Then, we have a valid information process (that is, no cheating), if $\mathcal{G}_t \subseteq \mathcal{G}_{t+1}$ (that is, it forms a filtration). More than an abstract mathematical concept, the field of stochastic programming uses this relationship to develop an explicit set of constraints to ensure a valid set of decisions (see Section 8.2.2).

5 Decisions

Dynamic systems evolve because of information that arrives from outside the system ("exogenous information processes") and decisions (which can be viewed as "endogenous information processes"). The challenge with decisions is determining how they are made, or how they should be made.

For our presentation, we focus purely on decisions that act on resources (this excludes, for example, pricing decisions). This means that all our decisions are constrained by resources, and their impact on the future is reflected in the resource constraints of future decisions.

5.1 Representing decisions

Although transportation problems are often viewed as consisting of decisions to move from one location to another, real problems also include a variety of other decisions, such as cleaning or repairing equipment, buying/selling equipment, sending crews home on rest, serving customers, loading/unloading equipment, reconfiguring equipment, moving equipment between customer-controlled equipment pools, and so on. To handle this generality we define

$$\begin{aligned}
\mathcal{C}^{\mathrm{D}} &= \text{set of decision classes (move to a location, clean/repair,} \\
&\qquad \text{serve a customer request, buy, sell),} \\
\mathcal{D}^c &= \text{set of decision types for class } c, \\
\mathcal{D} &= \bigcup_{c \in \mathcal{C}^{\mathrm{D}}} \mathcal{D}^c.
\end{aligned}$$

It is often necessary, for practical purposes, to define decision sets that depend on the attributes of the resource (which might be a layered resource). For this

purpose we let

$\mathcal{D}_a$ = set of decisions that can be used to act on a resource with attribute a.

An element $d \in \mathcal{D}$ is a type of decision. We measure the number of decisions made of each type using

x_{tad} = the number of resources of type a that a decision d is applied to using the information available at time t,

$x_t = (x_{tad})_{a\in\mathcal{A},d\in\mathcal{D}_a}$.

It is useful to contrast our indexing of decisions. It is common in deterministic models in transportation to use a variable such as x_{tij} to be the flow from i to j departing at time t. In a stochastic model, the outcome of a decision might be random, especially when we are working with vectors of attributes (for example, one attribute might indicate the maintenance status of a piece of equipment). The notation x_{tad} indexes the decision variable by information known when the decision is made, and not on the outcome (the notation for representing the outcome of a decision is given in Section 6).

5.2 The decision function

The real challenge, of course, is determining how to make a decision. We assume that we have a decision function that we represent using

$X_t^\pi(I_t)$ = a function that returns a vector x_t given information I_t,

I_t = the set of functions representing the information available at time t,

Π = a family of decision functions (or policies).

The decision function is required to return a feasible decision vector. We let

$\mathcal{X}_t$ = the set of feasible decision vectors given the information available at time t.

For example, we would expect $\mathcal{X}_t$ to include constraints such as

$$\sum_{d\in\mathcal{D}_a} x_{tad} = R_{t,ta}, \quad \forall a \in \mathcal{A}, \tag{7}$$

$$x_{tad} \leqslant u_{tad}, \quad \forall a \in \mathcal{A}, d \in \mathcal{D}_a, \tag{8}$$

$$x_{tad} \geqslant 0, \quad \forall a \in \mathcal{A}, d \in \mathcal{D}_a. \tag{9}$$

Of course, other constraints might arise in the context of a particular application. We note, however, that it is very common in deterministic models to treat as constraints serving a customer, possibly within a time window. These

are rarely expressed as constraints in dynamic models, simply because it is too easy for the model to be infeasible. We encourage limiting $\mathcal{X}_t$ to true physical constraints (such as flow conservation or hard constraints on flow), and let other "desirable behaviors" (such as serving a customer on time) be handled through the objective function.

The information I_t is all the data needed to make a decision, and only the data needed to make a decision. There is, of course, a close relationship between the information available and the type of decision function that is used. In our presentation, it is useful to identify four major classes of decision functions that reflect the type of information that is available. Each class corresponds with a well-established algorithmic strategy. The four classes and corresponding algorithmic strategies are summarized in Table 2. The first class uses what we know now, which is a standard myopic model. This covers the vast array of deterministic models which can be solved by algorithms developed for this problem class. This class, of course, is the foundation for the three other classes, which are all assumed to use this information (and these algorithmic strategies), augmented by the information listed for each class.

The second class includes forecasts of future information (this is not events in the future, but rather information that has not yet arrived). When we include information about the future, we can include a point forecast (which is the most widely used) or a distributional forecast. Point forecasts produce the rolling horizon procedures that are widely used in engineering practice (algorithmically, these can be solved using the same techniques as myopic models, but the problems are usually quite a bit larger). Alternatively, we can use a distributional forecast (where we explicitly represent multiple future scenarios) which puts us in the realm of stochastic programming. These are widely used in financial asset allocation, but have not been successfully applied to problems in transportation and logistics.

The third information class represents a forecast of the impact of a decision now on the future. This puts us into the framework of dynamic programming, where the future impact of a decision now is captured through a value function

Table 2.
Four classes of information and corresponding algorithmic strategies

	Information class	Algorithmic strategy
1	Knowledge: The data that describes what we know about the system now (myopic models)	Classical deterministic math programming
2a	Forecasts of exogenous information processes (point forecasts)	Rolling horizon procedures
2b	Forecasts of exogenous information processes (distributional forecasts)	Stochastic programming
3	Forecasts of the impacts now on the future (value functions)	Dynamic programming
4	Forecasts of future decisions (patterns)	Proximal point algorithms

(also known as a cost-to-go function). A value function can be viewed as the value of being in a particular state in the future, or for some classes, a forecast of a future dual variable.

The fourth and last information class is a forecast of a future decision. This can be accomplished in several ways, but a simple example is of the form "we normally put this type of locomotive on this type of train" or "we normally put our sleeper teams on long loads". Each of these examples summarizes a behavioral pattern for a company, represented in a form of an aggregated attribute–action pair (it is a type of decision acting on a resource with a subset of attributes).

It is possible, of course, to mix and match algorithmic strategies, but any reasonable strategy would start with the first information class and then seek to improve the solution by adding additional information classes.

6 System dynamics

We next have to model the impact of decisions on the system over time. We represent these effects using a device called the *modify function* which works at the level of an individual resource. The modify function can be represented as the mapping

$$M(t, a, d) \rightarrow (a', c, \tau). \tag{10}$$

Here, we are acting on a resource with attribute vector a with decision type d at time t (which also defines the information available when the decision is implemented). a' is the attribute of the resource after the decision, c is the contribution (or cost), and τ is the time required to complete the decision. It is useful to also write the results of the modify function as functions themselves

$$M(t, a, d) \rightarrow \big(a^{\mathrm{M}}(t, a, d), c^{\mathrm{M}}(t, a, d), \tau^{\mathrm{M}}(t, a, d)\big). \tag{11}$$

Here $a^{\mathrm{M}}(t, a, d)$ is the *attribute transition function*, where the superscript "M" is used to help identify the difference between the attribute vector a and the attribute transition function $a^{\mathrm{M}}(t, a, d)$. It is common to use $c_{tad} = c^{\mathrm{M}}(t, a, d)$ and $\tau_{tad} = \tau^{\mathrm{M}}(tad)$.

There are many problems where the evolution of the attribute vector is deterministic. This is particularly true if the only source of randomness is external customer demands. But many problems exhibit behaviors such as equipment failures or random delays due to weather or traffic. For these situations, we need to model the evolution of the attribute vector a_t just as we model the evolution of the state variable S_t. The attribute vector a_t in (11) would be a pre-decision attribute vector. We would change our attribute transition function so that it parallels our transition function, where we might write

$$a_{t+1} = a^{\mathrm{M}}(a_t, d_t, W_{t+1}),$$

where W_{t+1} is the exogenous information that describes equipment failures, delays, or the decision by a customer that a vehicle is dirty or unacceptable.

Any model that manages resources over time would have software that roughly corresponds to the modify function (which might be a single routine or several performing component tasks). The modify function can be arbitrary, rule-based logic, and as such is extremely flexible. However, this is very difficult to use when writing equations, so for this purpose we define

$$\delta_{t,t'a'}(t,a,d) = \begin{cases} 1, & \text{if } a^{\mathrm{M}}(t,a,d) = a', \\ 0, & \text{otherwise.} \end{cases}$$

The δ function is simply an indicator function that captures the outcome of a decision. We note that it is possible to capture uncertainty in the outcome of a decision. This is one reason why our decision variables x_{tad} are indexed with information known when a decision is made, and not the outcome of the decision. While this somewhat complicates tasks such as summing the flows into a location, it provides an important level of generality (for example, the time required to complete a decision might easily be random).

To model the resource dynamics, we are going to assume that only the vector x_{tt} is implementable (that is, if we are choosing a plan $x_{tt'}$, $t' > t$, these plans are ignored and replanned in the next time interval). We also assume that x_{tt} can only act on actionable resources (that is, those in the vector R_{tt}, as we did in Equation (7)). Our resource dynamics are given by

$$R_{t+1,t'a'} = R_{t,t'a'} + \sum_{a\in\mathcal{A}} \sum_{d\in\mathcal{D}_a} \delta_{t,t'a'}(t,a,d) x_{tad} + \widehat{R}_{t+1,t'a'}, \quad t' > t. \tag{12}$$

We note that the right-hand side of (12) does not include the case $t = t'$. Resources in the vector R_{tt} are all captured by the decision vector x_t (including the resources where we are doing nothing). Reading across, this equation says: the resources that we will know about at the end of time period $t+1$ which are actionable at time $t' \geqslant t+1$ consist of (a) the resources that are actionable at time t' that we knew about at time t (but could not act on), (b) the resources that were actionable at time t (we assume we act on everything, even if the action is to "do nothing"), and (c) the new resources entering the system during time interval $t+1$ which are actionable at time t'.

Although it is clearer to write out this equation in this form, it is more compact to write it out in matrix form. Let $\Delta_{tt'}$ be the matrix where $\delta_{t,t'a'}(t,a,d)$ is the element in row a', column (a,d). As before, $x_{tt'}$ is a vector with element (a,d). We can then rewrite (12) as

$$R_{t+1,t'} = R_{t,t'} + \Delta_{tt'} x_{tt'} + \widehat{R}_{t+1,t'}. \tag{13}$$

7 An optimization formulation

We are now ready to state our problem in the form of an optimization model. Our problem is to maximize the total contribution received over all time periods. Our contribution function in period t is given by

$$C_t(x_t) = \sum_{d\in\mathcal{D}} \sum_{a\in\mathcal{A}} c_{tad} x_{tad}.$$

We can write our decision model in the general form

$$X_t^\pi(R_t) = \arg\max_{x_t\in\mathcal{X}_t} C_t^\pi(x_t). \tag{14}$$

Note the difference between the actual contributions in a time period, $C_t(x_t)$, and the contributions that we are optimizing over to obtain a set of decisions, $C_t^\pi(x_t)$. Later, we provide specific examples of $C_t^\pi(x_t)$, but for now it suffices to establish the relationship between a policy and the specific contribution function being used to make a set of decisions. To establish our objective function, we first define

$$F_t^\pi(R_t) = E\left\{\sum_{t'=t}^{T} C_{t'}\left(X_{t'}^\pi\right)\Big|R_t\right\}.$$

$F_t^\pi(R_t)$ is the expected contribution from following policy π starting at time t in state R_t. The best policy is found by solving

$$F_0^*(\mathcal{S}_0) = \sup_\pi F_0^\pi(\mathcal{S}_0). \tag{15}$$

We note that we use the supremum operator to handle the mathematical possibility that we are searching over a class of policies that might not be bounded. For example, a policy might be characterized by a parameter that is required to be strictly greater than zero, but where zero produces the best answer. Or, the optimal parameter could be infinite. If the set of policies is contained in a compact set (that is, a set that is closed and bounded), then we can replace the supremum operator by a simple maximization operator.

Equation (15) is computationally intractable, so we generally find ourselves looking at different classes of policies, and finding the best within a class. If this were a deterministic problem (that is, no new information arriving over time), we could write it in the following form

$$\max_x \sum_{t\in\mathcal{T}} c_t x_t \tag{16}$$

subject to

$$A_t x_t - \sum_{t''<t} \Delta_{t''t} x_{t''} = \widehat{R}_t, \tag{17}$$

$$x_t \geqslant 0. \tag{18}$$

In this formulation, the decision variable is the vector of decisions x_t. When the problem is stochastic, it becomes a lot harder to write down the decision variables, because the decisions we make will depend on the state of the system at time t or, more generally, on the particular scenario. For this reason, we cannot simply maximize the expectation, because the expectation is summing over many scenarios, and we can make different decisions for each scenario (this would not be the case if the decision is a static parameter such as the capacity of a facility or the size of the fleet of vehicles).

An alternative approach is to assume that instead of choosing all the decisions, over all the time periods, over all the scenarios, that we are going to find a single decision function $X_t^\pi(I_t)$ that returns a decision given the information I_t. The challenge is that this single function has to work well across all the scenarios (which produces different information sets). However, if the function is defined to work only with the information available at time t, then it will automatically satisfy the various measurability/nonanticipativity conditions that we require. Most people will generally find that this approach is more natural.

When we use this approach, our optimization problem looks like

$$\max_{\pi \in \Pi} E\left\{ \sum_{t \in \mathcal{T}^{ph}} c_t X_t^\pi(I_t) \right\}, \tag{19}$$

where $\mathcal{T}^{ph}$ is the set of time periods that define our planning horizon (say, $t = 0, \ldots, T$). Unlike classical optimization, we do not specify all the constraints, since we already require that the decision function produce decisions that satisfy the constraints $\mathcal{X}_t$ at a point in time. As a result, we have only to specify constraints that govern the evolution of the system over time, as with Equation (12).

We are not going to attempt to solve Equation (19) exactly. We quickly give up the search for optimal solutions (or even optimal policies) if we can find "good" policies that work well over as many scenarios as possible. Such policies are often referred to as *robust*. In the next section, we address the problem of finding good policies.

8 Algorithmic strategies

Unlike deterministic optimization problems where there is a well-defined concept of an optimal solution (in the case of integer programs, these can be difficult to both find and identify, but at least they are well defined). In the dynamic world, there will be valid disagreements over what even constitutes a valid policy.

In this section, we summarize four major classes of policies, and offer these as choices for modelers to consider and adapt to their needs. These are complementary strategies, since they can be used in various combinations depending on the characteristics of the problems at hand. Each policy is going to use a

new class of information, and open up a new algorithmic strategy. But all the strategies start with the first strategy.

8.1 Myopic models

The most basic strategy on which all other methods build is a myopic policy, which optimizes over information that is known now, but extends into the future (we might take a driver available three hours from now and assign him to a load that will be available five hours from now). The resulting model can be stated as

$$\max_{x_0} \sum_{t' \in \mathcal{T}^{ph}} c_{0t'} x_{0t'} \tag{20}$$

subject to

$$A_{00} x_{00} = R_{00}, \tag{21}$$

$$A_{0t'} x_{0t'} - \sum_{t''=0}^{t'-1} \Delta_{t''t'} x_{0t''} = \widehat{R}_{0t'}, \quad t' \geqslant 1, \tag{22}$$

$$x_{0t'} \geqslant 0 \text{ and possibly integer.} \tag{23}$$

Here, $x_{0t''}$ is a vector of decisions to be implemented at time t'' using the information available at time 0. The matrix $\Delta_{t''t'}$ is an indicator matrix that tells us when a decision implemented at time t'' produces a resource that is actionable at time t'. Technically, we need a third time index to show that $\Delta_{t''t'}$ is computed with information available at time 0.

With this generic model we include the entire range of classical optimization problems, and algorithms that have been studied in transportation and logistics. These might be simple assignment problems, vehicle routing problems, network design problems and multicommodity flow problems. In general, however, these are much smaller than the types of problems that arise in planning applications. We assume that this basic myopic problem is solvable to an acceptable level of optimality. This is the fundamental building block for all other modeling and algorithmic strategies.

There are some problems where it is possible to provide effective solutions using a simple myopic model. Assigning taxi drivers to customers, assigning long-haul truckload drivers to loads, and the dynamic routing and scheduling of pickup and delivery trucks in a city. But our experience has been that while these models can be effective, they are always missing something.

8.2 Rolling horizon procedures

We can extend our basic myopic model by including information that is forecasted to come in the future. A common procedure is to use a point forecast in a rolling horizon procedure (Section 8.2.1), but we can also use a distributional forecast in a stochastic programming algorithm (Section 8.2.2).

8.2.1 Deterministic models

The most common approach to combining current knowledge with forecasted information is to use a point forecast of future events which produces a classic rolling horizon procedure using point forecasts. Let $\overline{R}_{tt'}$ be the forecast of information that becomes available during time interval t that is actionable at time t'. We note that these forecasts are known at time 0. Using these forecasts, our optimization problem becomes

$$\max_{x_0}\left(\sum_{t\in\mathcal{T}^{ph}}(c_{0t}x_{0t})+\sum_{t'\in\mathcal{T}^{ph}\setminus 0}c_{t'}x_{0t'}\right) \tag{24}$$

subject to

$$A_{00}x_{00}=R_{00}, \tag{25}$$

$$A_{0t}x_{0t}-\sum_{t'=0}^{t-1}\Delta_{t't}x_{0t'}=\sum_{t'=0}^{t}\overline{R}_{t't},\quad t>0, \tag{26}$$

$$x_{0t}\geqslant 0\quad \text{for } t\in\mathcal{T}^{ph}. \tag{27}$$

Note that the right-hand side of Equation (26) is the total number of resources that are actionable at time t'. To obtain this, we have to sum over all the information that is forecasted to arrive in future time periods. It is interesting that many practitioners view this as the "correct" way to solve dynamic problems. They recognize that point forecasts are not perfect, but argue that you simply reoptimize as new information comes in.

Rolling horizon procedures have the primary benefit that they use classical modeling and algorithmic technologies. In addition, point forecasts are the easiest to understand in the business community. But as with myopic models, they will over time start to show their weaknesses. For example, rail customers often exhibit highly variable demands. A forecasting system will tend to forecast the average demand, while a seasoned operations manager will know that you need to supply the customer more than the average to cover the days when the demand is high. Waiting until the last minute might not provide enough time to move cars to satisfy the demand.

Another problem class where point forecasts perform poorly are discrete routing and scheduling. Here, a customer demand will normally be zero but will sometimes be one. A point forecast would produce a fraction such as 0.2. It is not possible to route an entire vehicle to each customer with a nonzero (fractional) demand forecast. Randomly sampling the demands is nothing more than routing to a random outcome. Allowing a model to send a fraction of a vehicle to a fraction of a demand will create an extremely difficult integer program for the first period.

8.2.2 Stochastic programming

If we wish to explicitly model the possibility of multiple outcomes, we find ourselves in the domain of stochastic programming. Now, we have to allow for

the possibility that different decisions will be made in the future depending on the actual evolution of information.

In this case, our optimization problem becomes

$$\max_{x}\left(\sum_{t\in\mathcal{T}^{ph}}(c_{0t}x_{0t})+\sum_{\omega\in\widehat{\Omega}}\hat{p}(\omega)\sum_{t\in\mathcal{T}^{ph}\setminus 0}\sum_{t'>t}c_{tt'}x_{tt'}(\omega)\right) \tag{28}$$

subject to, for $t, t' \in \mathcal{T}^{ph}$:

(1) First stage constraints:

$$A_{00}x_{00} = R_{00}, \tag{29}$$

$$A_{0t'}x_{0t'} - \sum_{t''=0}^{t'-1}\Delta_{t''t'}x_{0t''} = R_{0t'}, \quad t' > 0, \tag{30}$$

$$x_{0t'} \geqslant 0, \quad t' \geqslant t. \tag{31}$$

(2) Later stage constraints, for all $\omega \in \widehat{\Omega}$, $t, t' \in \mathcal{T}^{ph} \setminus 0$, $t' \geqslant t$:

$$A_{tt}(\omega)x_{tt}(\omega) = R_{tt}(\omega), \tag{32}$$

$$A_{tt'}(\omega)x_{tt'}(\omega) - \sum_{t''=t}^{t'-1}\Delta_{t''t'}(\omega)x_{tt''}(\omega) = \widehat{R}_{tt'}(\omega), \tag{33}$$

$$x_{tt'}(\omega) \geqslant 0. \tag{34}$$

This formulation requires that we choose a vector of decisions for the first time period x_{0t} but allows a different set of decisions for each scenario in the later time periods. This might not be a bad approximation, but it does allow the decisions in later time periods to "see" future information. This happens because we are indexing a decision by an outcome ω which implicitly determines all future information. In the vocabulary of stochastic programming, we prevent this by adding what are known as nonanticipativity constraints. We can express these by using some of our notation (and concepts) from Section 4.4. As before, let $h_t = (\omega_1, \omega_2, \ldots, \omega_t)$ be the history of the process up through time t, and let $\mathcal{H}_t$ be the set of all possible histories. In addition, define the subset $\Omega_t(h_t) = \{\omega \in \widehat{\Omega} \mid (\omega_1, \omega_2, \ldots, \omega_t) = h_t\}$ to be the set of scenarios $\omega \in \widehat{\Omega}$ which match the history h_t up through time t. We would then add the constraint

$$x_t(h_t) = x_t(\omega), \quad \forall \omega \in \Omega_t(h_t), h_t \in \mathcal{H}_t. \tag{35}$$

Equation (35) is our nonanticipativity constraint, which has the effect of forcing the set of decisions $(x(\omega), \omega \in \widehat{\Omega})$ to form a filtration. Literally, we want each set of decisions with the same history to be the same.

For most transportation problems, the optimization problem (28)–(35) is too large to be solved, although special cases exist (see, for example, Morton et al., 2003; Glockner and Nemhauser, 2000).

8.3 Dynamic programming

A powerful technique for taking into account uncertain future events is dynamic programming, which uses the construct of a value function to capture the expected value of being in a particular state in the future. If we can let R_t represent the state of our system, we can calculate the value $V_t(R_t)$ of being in state R_t from time t onward using the well-known Bellman equations

$$V_t(R_t) = \max_{x_t \in \mathcal{X}_t} \left(C_t(x_t) + \sum_{R' \in \mathcal{R}} P(R'|R_t, x_t) V_t(R') \right), \tag{36}$$

where $P(R'|R_t, x_t)$ is the probability of being in state R' if we are in state R_t and take action x_t. A more general way of writing Equation (36) (and more useful for our purposes here) is using

$$V_t(R_t) = \max_{x_t \in \mathcal{X}_t} \big[C_t(x_t) + E\big\{ V_{t+1}(R_{t+1}) | R_t \big\} \big], \tag{37}$$

where $R_{t+1} = R_t + \Delta_t x_t + \widehat{R}_{t+1}$. It is generally assumed that we are going to compute Equation (36) for each possible value of R_t. As we computed in Equation (1), the number of possible values of R_t is extremely large even for fairly small problems. This is the well-known "curse of dimensionality" that has been recognized since dynamic programming was first discovered. For our problem class, there are actually three curses of dimensionality: the state space, the outcome space, and the action space. Since new information (such as $\widehat{R}_t$) is typically a vector, sometimes of high dimensionality, the expectation in Equation (37) will generally be computationally intractable. The more conventional form of Bellman's equation, shown in Equation (36) hides this problem since the one-step transition matrix is itself an expectation. Finally, if we compute $V_t(R_t)$ for each discrete value of R_t, then the only way to find the optimal x_t is to search over each (discrete) element of the set $\mathcal{X}_t$. Again, if x_t is a vector (for our problems, the dimensionality of x_t can easily be in the thousands or tens of thousands) the size of $\mathcal{X}_t$ is effectively infinite.

The dynamic programs depicted in Equations (36) and (37) assume a particular representation of states, information and action. We can write the *history* of our process up through time t as

$$h_t = (R_0, x_0, \omega_1, R_1, x_1, \ldots, \omega_t, R_t, x_t). \tag{38}$$

The evolution of the state variable is often depicted using

$$R_{t+1} = R^{\mathrm{M}}(R_t, x_t, \omega_{t+1}), \tag{39}$$

where $R^{\mathrm{M}}(\cdot)$ captures the dynamics of the system. We can write the decision function as

$$X_t^{\pi}(R_t) = \arg\max_{x_t \in \mathcal{X}_t} \big(c_t x_t + E\big(V_{t+1}(R_{t+1}) | R_t \big) \big), \tag{40}$$

Since the expectation is typically computationally intractable, we can replace this with an approximation. Let $\widehat{\Omega} \subset \Omega$ be a small set of potential outcomes. This would be written

$$X_t^{\pi}(R_t) = \arg\max_{x_t \in \mathcal{X}_t} \left(c_t x_t + \sum_{\hat{\omega} \in \widehat{\Omega}} \hat{p}(\hat{\omega}) V_{t+1}\big(R_{t+1}(\hat{\omega})\big) \right), \tag{41}$$

where $\hat{p}(\hat{\omega})$ is the probability of the sampled outcome $\hat{\omega} \in \widehat{\Omega}$. The problem we encounter in practice is that even when $\widehat{\Omega}$ is relatively small, it can still greatly complicate solving (41). The problem we encounter in transportation and logistics is that the myopic problem

$$X_t^{\pi}(R_t) = \arg\max_{x_t \in \mathcal{X}_t} c_t x_t \tag{42}$$

can be computationally difficult. This could be a vehicle routing problem or integer multicommodity flow problem. Adding the summation over multiple outcomes (as we do in Equation (41)) can make a hard problem much larger and harder. We could avoid this by using a single sample $\omega_{t+1} = W_{t+1}(\omega)$, producing

$$X_t^{\pi}(R_t, \omega) = \arg\max_{x_t \in \mathcal{X}_t} \big(c_t x_t + V_{t+1}\big(R_{t+1}(\omega_{t+1})\big) \big). \tag{43}$$

Now we are choosing x_t given $R_{t+1}(\omega_{t+1})$ which is like making a decision now given the information that will arrive in the next time period. This will not even be a good approximation.

We can overcome this problem by introducing the concept of measuring the state of the system immediately *after* a decision is made (or equivalently, before new information has arrived). We refer to our original state variable R_t as the "pre-decision" state variable, and we introduce a new state variable, denoted R_t^x as the "post-decision" state variable. We can now write the history of our process as

$$h_t = \big(R_0, x_0, R_0^x, \omega_1, R_1, x_1, R_1^x, \ldots, \omega_t, R_t, x_t, R_t^x\big). \tag{44}$$

If we write the evolution of our system in terms of R_t^x, we obtain

$$R_t^x = R_t^{\mathrm{M}}\big(R_{t-1}^x, \omega_t, x_t\big). \tag{45}$$

Writing the Bellman equations around R_t^x we obtain

$$V_{t-1}^x\big(R_{t-1}^x\big) = E\Big\{ \max_{x_t \in \mathcal{X}_t} \big(c_t x_t + V_t^x\big(R_t^x(x_t)\big) \big) \big| R_{t-1}^x \Big\}. \tag{46}$$

The relationship between the pre- and post-decision value functions is best illustrated using

$$V_t(R_t) = \max_{x_t \in \mathcal{X}_t} c_t x_t + V_t^x\big(R_t^x(x_t)\big),$$
$$V_{t-1}^x\big(R_{t-1}^x\big) = E\big\{ V_t(R_t) | R_{t-1}^x \big\}.$$

Using Equation (46) we still encounter the problem of computing the expectation, but now consider what happens when we drop the expectation and solve it for a single sample realization ω

$$X_t^\pi(R_{t-1}, \omega_t) = \arg \max_{x_t \in \mathcal{X}_t(\omega)} c_t x_t + V_t^x\big(R_t^x(x_t)\big). \tag{47}$$

Here, we are finding the decision x_t given the information $\omega_t = \widehat{R}_t(\omega)$ which would be available when we make a decision. As a result, we are not violating any informational (measurability or anticipativity) constraints. What is most important is that $V_t^x(R_t^x(x_t))$ is a deterministic function of x_t, so we do not have to work with even an approximation of an expectation (the approximation is implicit in $V_t^x(R_t^x(x_t))$).

We have now crossed a major hurdle, but we still have the problem that we do not know $V_t^x(R_t^x)$. We are going to need to use an approximation which will be estimated iteratively. At iteration n, we would solve (47) using the approximation from iteration $n-1$

$$X_t^\pi\big(R_{t-1}, \omega_t^n\big) = \arg \max_{x_t \in \mathcal{X}_t(\omega^n)} c_t x_t + \overline{V}_t^{n-1}\big(R_t^x(x_t)\big), \tag{48}$$

where we drop the superscript x when writing the approximation $\overline{V}$. We denote the function produced by solving the right-hand side of (48) by

$$\widetilde{V}_t^n\big(R_t^n(\omega^n)\big) = \max_{x_t \in \mathcal{X}_t(\omega^n)} c_t x_t + \overline{V}_t^{n-1}\big(R_t^x(x_t)\big). \tag{49}$$

A brief note on notation is in order here. $\widetilde{V}_t^n(R_t(\omega^n))$ is an approximation of $V_{t-1}^x(R_{t-1}^x)$, which suggests an inconsistency in how we are indexing by time (V^x is indexed by t, while $\widetilde{V}$ is indexed by $t-1$). Recall that our time indexing indicates the information content. V_{t-1}^x is an expectation conditioned on R_{t-1}^x, and therefore contains information up through $t-1$. $\widetilde{V}_t^n(R_t(\omega^n))$ is a conditional approximation, computed using the information that arrives in time t. It does not represent even an approximation of an expectation (by contrast $\overline{V}_t^n(R_t)$ is an approximation of the expected value function conditioned on R_t).

We use the information from solving (49) to update the value function approximation. How this update is performed depends on the type of approximation used, so for now we represent this updating process simply as

$$\overline{V}_{t-1}^n = U^V\big(\overline{V}_{t-1}^{n-1}, \widetilde{V}_t^n, R_t^n(\omega^n)\big). \tag{50}$$

The algorithm works by stepping forward through time using the value function approximation from the previous iteration. This is the standard approach used in the field of approximate dynamic programming. Instead of finding the value function over all possible states, we develop approximations around the states that are actually visited. The two textbooks that describe these tech-

niques (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998) both describe these procedures in the context of steady state problems. Operational problems in transportation and logistics require time-dependent formulations.

There are two strategies for updating the value function (especially when a time-dependent formulation is used). The first uses a single, forward pass, updating the value function approximations as we proceed. The steps of this procedure are outlined in Figure 3. The second uses a two-pass procedure, stepping forward through time determining states and actions, and a backward pass that updates the value function using information about the states that were actually visited. The steps of this procedure are given in Figure 4.

In our applications, we will generally be more interested in derivatives than the actual value function itself. When we solve the approximation $\widetilde{V}_t(R_t)$ in Equation (49), we will typically be solving these problems subject to flow conservation constraints on the resources as in Equation (7). As a general rule our

Step 0. Initialization:
  Initialize $\overline{V}_t^0$, $t \in \mathcal{T}$.
  Set $n = 1$.
  Initialize R^1.
Step 1. Do while $n \leqslant N$:
  Choose $\omega = (\omega_1^n, \omega_2^n, \ldots, \omega_T^n)$.
Step 2. Do for $t = 1, 2, \ldots, T$:
  Step 2a. Solve Equation (49) to obtain $\widetilde{V}_t^n(R_t^n)$ and x_t^n.
  Step 2b. Compute $R_t^{x,n} = R_t^{\mathrm{M}}(R_{t-1}^{x,n}, \omega_t^n, x_t^n)$.
  Step 2c. Update the value function approximations using $\overline{V}_{t-1}^n \leftarrow U^V(\overline{V}_{t-1}^{n-1}, \widetilde{V}_t^n, R_{t-1}^{x,n})$.
Step 3. Return the policy $X_t^\pi(R_t, \overline{V}_t^N)$.

Fig. 3. Single pass version of the approximate dynamic programming algorithm.

Step 0. Initialize $\overline{V}_t^0$, $t \in \mathcal{T}$.
  Set $n = 1$.
  Initialize R^1.
Step 1. Do while $n \leqslant N$:
  Choose $\omega = (\omega_1^n, \omega_2^n, \ldots, \omega_T^n)$.
Step 2. Do for $t = 0, 1, \ldots, T-1$:
  Step 2a. Solve Equation (49) to obtain $\widehat{V}_t^n(R_t)$ and x_t^n.
  Step 2b. Compute $R_t^{x,n} = R_t^{\mathrm{M}}(R_{t-1}^{x,n}, \omega_t^n, x_t^n)$.
Step 3. Do for $t = T-1, T-2, \ldots, 1$:
  Step 3a. Recompute $\widetilde{V}_t^n(R_t^n)$ using $\overline{V}_{t+1}^n$ and the decision x_t^n from the forward pass:
$$\widetilde{V}_t^n(R_t^n) = c_t x_t + \widetilde{V}_{t+1}^n(R_{t+1}^n). \tag{51}$$
  Step 3b. Update the value function approximations, $\overline{V}_{t-1}^n \leftarrow U^V(\overline{V}_{t-1}^{n-1}, \widetilde{V}_t^n, R_{t-1}^{x,n})$.
Step 4. Return policy $X_t^\pi(R_t, \overline{V}^N)$.

Fig. 4. Double pass version of the approximate dynamic programming algorithm.

problems are continuous but nondifferentiable in R_t. If we are solving linear resource allocation problems, we can obtain stochastic subgradients just by using the dual variable for the flow conservation constraint. In general, these dual variables are subgradients of $\widetilde{V}_t^n(R_t)$ only, which depends on $\overline{V}_t^{n-1}(R_t(x_t))$. These can be effective, but the use of approximations from previous iterations will slow the overall convergence.

Ideally, we would prefer to have gradients of all future profits. Let

$$V_t^{\pi}(R_t, \omega) = \sum_{t'=t}^{T} c_{t'} X_{t'}^{\pi}(R_{t'}, \omega) \tag{52}$$

be the total future profits given an initial resource vector R_t, and sample realization ω, under policy π. We can obtain the gradient of $V_t^{\pi}(R_t, \omega)$ by storing the basis obtained as we step forward in time. The basis will allow us to identify the effect of, say, adding one more unit to an element R_{ta} by giving us the impact on costs in time period t, and the impact on future resources R_{t+1}. By storing this information as we step forward in time, we can then compute the effect of increasing R_{ta} by one by stepping backward through time, combining the impact of an additional unit of resource of each type at each time $t' \geqslant t$ with the value of one more unit of resource at the next time period.

8.4 Incorporating low-dimensional patterns

In practice, it is common to carefully develop a math programming-based optimization model, only to find that a knowledgeable expert will criticize the behavior of the model. It is possible, although generally unlikely, that the expert is simply finding a better solution than the algorithm. More often, the expert simply recognizes behavior that will incur a cost that is not captured by the model, or expresses a preference for an alternative behavior because of other unquantified benefits.

One response to such input is to "fix the model" which means improving the cost function or adding constraints that capture the physical problem more realistically. For some problems, this is the only solution. However, it can often take weeks or months to implement these fixes, which might easily require data that was not originally available. Often, the complaints about the model can be expressed in fairly simple ways:

- You should avoid sending that type of cargo aircraft into Saudi Arabia since we do not have the ability to handle certain types of repairs.
- Do not send Cleveland-based drivers into Indianapolis because their union is in a different conference and it creates a lot of bookkeeping problems.
- You should put your sleeper teams on longer loads so that you get higher utilization out of the two drivers.

- Do not use the six-axle locomotives on certain trains out of Denver because the track is too curved to handle the long locomotives.
- Try to send flatcars that can handle trailers into Chicago because that location moves a lot of trailers from trucking companies.
- Avoid sending drivers domiciled in Florida into Texas because we do not have very much freight out of that location that terminates in Florida, making it hard to get drivers home.
- Avoid putting inexperienced drivers on service-sensitive loads.

All of these preferences can be stated as low dimensional patterns, each of which consists of three elements: the attributes of the resource being acted on at some level of aggregation, the decision being made (also at some level of aggregation), and the fraction of time that the pattern should occur. Aggregation on the attributes arises because each rule is typically trying to accomplish a particular objective, such as getting a driver home or avoiding sending equipment needing maintenance into areas that do not have the proper repair facilities. A pattern measures how often a particular decision is applied to a group of resources, and the decisions will also generally be expressed at an aggregate level. For example, we are not specifying a rule that a type of driver should (or should not) be assigned to move a particular load of freight. Instead, we are expressing preferences about types of loads (high priority, loads into a particular destination). Finally, these "rules" are generally not hard constraints, but rather expressions of behaviors to be encouraged or discouraged.

Each pattern can be thought of as a preference about the desirability of a particular type of state action pair, which can be represented as aggregations on the attribute vector and decision. For this, we define

$$\begin{aligned}
\mathcal{P} &= \text{a collection of different patterns,}\\
G_a^p &= \text{aggregation function for pattern } p \in \mathcal{P} \text{ that is applied}\\
&\qquad \text{to the attribute vector } a,\\
G_d^p &= \text{aggregation function for pattern } p \in \mathcal{P} \text{ that is applied}\\
&\qquad \text{to the decision } d.
\end{aligned}$$

These aggregation functions create new attribute spaces and decision sets

$$\begin{aligned}
\mathcal{A}^p &= \left\{G_a^p(a) \mid a \in \mathcal{A}\right\},\\
\mathcal{D}_a^p &= \left\{G_d^p(d) \mid d \in \mathcal{D}_a\right\}.
\end{aligned}$$

For notational compactness, it is useful to let $\bar{a}$ represent a generic aggregated attribute vector, and $\bar{d}$ an aggregated decision. We use this notation when the level of aggregation is either not relevant, or apparent from the context. A pair $(\bar{a}, \bar{d})$ represents an aggregated state/action pair (literally, an attribute vector/action pair). The last element of a pattern is the fraction of time that we expect to observe decision $\bar{d}$ when we are acting on a resource with attribute $\bar{a}$,

which we represent as follows:

$$\rho^p_{\bar{a}\bar{d}} = \text{the fraction of time that resources with attributes } \{a \mid G^p_a(a) = \bar{a}\} \text{ are acted on using decisions } \{d \mid G^p_d(d) = \bar{d}\}.$$

Given a flow vector $x_t = (x_{tad})_{a\in\mathcal{A},d\in\mathcal{D}}$ returned by the model, the number of times that the flow matches a pattern is given by

$$\begin{aligned}\bar{x}^p_{t\bar{a}\bar{d}} &= G^p(x)\\ &= \sum_{\forall a\in\mathcal{A}} \sum_{\forall d\in\mathcal{D}_a} x_{tad} I_{\{G^p_a(a)=\bar{a}\}} I_{\{G^p_d(d)=\bar{d}\}},\end{aligned} \tag{53}$$

where $I_X = 1$ if X is true. We also need to express the resource state variable R_t in an aggregated form

$$\begin{aligned}\overline{R}^p_{t\bar{a}} &= \sum_{\forall a\in\mathcal{A}} R_{ta} I_{\{G^p_a(a)=\bar{a}\}}\\ &= \text{the number of resources with attribute } \bar{a} \text{ at time } t,\\ R_{\bar{a}}(x) &= \sum_{t\in\mathcal{T}} \sum_{\bar{d}\in\mathcal{D}^p_{\bar{a}}} x_{t\bar{a}\bar{d}}.\end{aligned}$$

Finally, we define the model pattern flows as

$$\begin{aligned}\rho_{t\bar{a}\bar{d}}(x) &= \frac{x_{t\bar{a}\bar{d}}}{R_{t\bar{a}}} \quad (\forall \bar{a}\in\mathcal{A}^p, \forall \bar{d}\in\mathcal{D}^p_{\bar{a}}, \forall t\in\{1,2,\ldots,T\})\\ &= \text{the fraction of time that resources with attribute } \bar{a} \text{ are acted on with decision } \bar{d} \text{ at time } t.\\ \rho_{\bar{a}\bar{d}}(x) &= \frac{\sum_{t=1}^{T} x_{t\bar{a}\bar{d}}}{R_{\bar{a}}} \quad (\forall \bar{a}\in\mathcal{A}^p, \forall \bar{d}\in\mathcal{D}^p_a)\\ &= \text{the fraction of time that resources with attribute } \bar{a} \text{ are acted on with decision } \bar{d} \text{ over the entire horizon.}\end{aligned}$$

The static flow fraction after iteration n can also be written as

$$\begin{aligned}\rho^n_{\bar{a}\bar{d}} &= \frac{\sum_{t\in\mathcal{T}} x^n_{t\bar{a}\bar{d}}}{R^n_{\bar{a}}}\\ &= \sum_{t\in\mathcal{T}} \frac{x^n_{t\bar{a}\bar{d}}}{R^n_{t\bar{a}}} \frac{R^n_{t\bar{a}}}{R^n_{\bar{a}}}\\ &= \sum_{t\in\mathcal{T}} \left(\rho^n_{t\bar{a}\bar{d}} \frac{R^n_{t\bar{a}}}{R^n_{\bar{a}}}\right).\end{aligned} \tag{54}$$

Our goal is to create a solution $\bar{x}^p$ that reasonably matches the exogenous pattern ρ^p. We do this by introducing the metric

$$H\big(\rho(x), \rho\big) = \sum_{p\in\mathcal{P}} \sum_{\bar{a}\in\mathcal{A}^p} \sum_{\bar{d}\in\mathcal{D}_a^p} R_{\bar{a}\bar{d}}\big(\rho_{\bar{a}\bar{d}}(x) - \rho^p_{\bar{a}\bar{d}}\big)^2. \tag{55}$$

We often refer to $H(\rho(x), \rho)$ as the "happiness function" since the domain expert evaluating the model tends to be "happier" when the model flows reasonably match the exogenous pattern. More formally, $H(\rho(x), \rho)$ is a well-defined distance matrix used in proximal point algorithms to produce solutions. We now wish to solve

$$\max_{\pi\in\mathcal{P}} \sum_{t\in\mathcal{T}} \big(c_t X_t^\pi - \theta H\big(\rho(x), \rho\big)\big), \tag{56}$$

where θ is a scaling parameter (these might depend on the specific pattern). Our decision function at time t is given by

$$X_t^\pi = \arg\max_{x_t\in\mathcal{X}_t} \big(c_t x_t - \theta H\big(\rho(x), \rho\big)\big).$$

We are generally unable to solve this problem, so we use instead sequences of approximations of the pattern metric, which for the moment we can write as

$$X_t^\pi = \arg\max_{x_t\in\mathcal{X}_t} \big(c_t x_t - \theta \widehat{H}^{n-1}\big(\rho^{n-1}(x), \rho\big)\big).$$

We encounter an algorithmic challenge because the proximal term is defined over time but we have to solve the problem one time period at a time. We overcome this problem by solving a series of separable, quadratic approximations of the pattern metric. We start by computing the gradient of the pattern metric, which we denote $\widehat{H}^n(x, \rho)$, by differentiating (55)

$$\begin{aligned}\widehat{H}^n_{\bar{a}\bar{d}}(\rho) &= \left.\frac{\partial H}{\partial \rho_{\bar{a}\bar{d}}}\right|_{\rho_{\bar{a}\bar{d}}=\rho^n_{\bar{a}\bar{d}}} \\ &= 2R^n_{\bar{a}}\big(\rho^n_{\bar{a}\bar{d}} - \rho_{\bar{a}\bar{d}}\big), \quad \forall \bar{a}\in\mathcal{A}^p, \forall \bar{d}\in\mathcal{D}^p_{\bar{a}},\end{aligned} \tag{57}$$

where p is the pattern associated with the level of aggregation in $\bar{a}$ and $\bar{d}$.

A convergent algorithm can be designed using a Gauss–Seidel strategy which optimizes the pattern metric computed by using flows $x^n_{t'}$ for $t' < t$, and $x^{n-1}_{t'}$ for $t' > t$, and then optimizing x_t. We then solve a separable, quadratic programming approximation at each point in time.

To describe the algorithm in detail, we first define

$$R^n_{\bar{a}}(t) = \sum_{t'=1}^{t-1} \sum_{\bar{d}\in\mathcal{D}^p_{\bar{a}}} x^n_{t'\bar{a}\bar{d}} + \sum_{t'=t}^{T} \sum_{\bar{d}\in\mathcal{D}^p_{\bar{a}}} x^{n-1}_{t'\bar{a}\bar{d}}$$

and

$$\tilde{\rho}^n_{\bar{a}\bar{d}}(t) = \sum_{t'=1}^{t-1}\left(\frac{x^n_{t'\bar{a}\bar{d}}}{R^n_{\bar{a}}(t')}\right) + \sum_{t'=t}^{T}\left(\frac{x^{n-1}_{t'\bar{a}\bar{d}}}{R^n_{\bar{a}}(t')}\right),$$
$$\forall \bar{a} \in \mathcal{A}, \forall \bar{d} \in \mathcal{D}^p_{\bar{a}}, \forall t \in \{1, 2, \ldots, T\}. \tag{58}$$

The Gauss–Seidel version of the gradient is now

$$\tilde{h}^n_{t\bar{a}\bar{d}} = 2R^n_{\bar{a}}(t)\big(\tilde{\rho}^n_{\bar{a}\bar{d}}(t) - \rho_{\bar{a}\bar{d}}\big), \quad \forall \bar{a} \in \mathcal{A}, \forall \bar{d} \in \mathcal{D}^p_{\bar{a}}, \forall t \in \{1, 2, \ldots, T\}. \tag{59}$$

Our algorithm proceeds by iteratively solving subproblems of the form

$$x^n_t = \arg\max \sum_{a\in\mathcal{A}} \sum_{d\in\mathcal{D}_a} c_{tad}x_{tad} - \theta \sum_{p\in\mathcal{P}} \sum_{\bar{a}\in\mathcal{A}^p} \sum_{\bar{d}\in\mathcal{D}^p_{\bar{a}}} \left(\frac{1}{R^n_{\bar{a}}(t)} \sum_{\bar{d}\in\mathcal{D}^p_{\bar{a}}} \int_0^{x_{t\bar{a}\bar{d}}} \big[\tilde{h}^n_{t\bar{a}\bar{d}} + 2\big(u - x^{n-1}_{t\bar{a}\bar{d}}\big)\big]\,\mathrm{d}u\right). \tag{60}$$

This approach has been found to produce solutions that closely match exogenous patterns within as few as two or three iterations.

The real value of exogenous patterns is that it allows an expert to change the behavior of the model quickly by editing an external data file. This approach does not replace the need to ensure the best quality model through careful engineering, but it does allow us to produce more realistic behaviors quickly by simply changing an external data file.

8.5 Summary: The four information classes

This section has outlined a series of modeling and algorithmic strategies that can be largely differentiated based on the information being made available to the model. We started in Section 8.1 by providing a basic myopic model that only uses the *information* available at time t. There was no attempt to incorporate any information that had not yet arrived. All the remaining models use some sort of forecast. Section 8.2 illustrates the use of point and distributional forecasts of future information, which produces classical rolling horizon procedures. Section 8.3 introduces a forecast of the value of resources in the future, represented as a value function. This information brings into play the theory of dynamic programming. Finally, Section 8.4 uses what might be called a forecast of a decision, expressed in the form of low-dimensional patterns. As an alternative, we could have avoided mathematical optimization entirely and resorted instead to an extensive set of rules; this would be the approach that the artificial intelligence or simulation communities would have adopted. For transportation problems, the rules simply become too complex. Instead, we have found that expert knowledge can often be represented in the form of low

dimensional patterns, and incorporated using a proximal point term, bringing into play the theory of proximal point algorithms.

Up to now, we have represented our decision functions as depending on the resource vector R_t. In practice, what we know at time t can also include other data such as estimates of system parameters such as travel times and costs, and functions such as forecasting models. It is useful to summarize "what we know now" as consisting of data knowledge (such as the resource vector) and functional knowledge (such as models for forecasting demand). Let ν_t be our estimates at time t of various parameters, such as those that govern the modify function. Our data knowledge at time t is then given by

$$K_t = (R_t, \nu_t).$$

We differentiate between a forecast model (which we know now) from an actual forecast of an event in the future. A forecast is an actual generation of information that might arrive in the future. Let

$$\begin{aligned}
\omega_{[t,T^{ph}]} &= \text{a vector of potential information events over} \\
&\qquad \text{the planning horizon } \left(t, t+1, \ldots, T^{ph}\right), \\
\Omega_{[t,T^{ph}]} &= \text{the set of all potential information events that we wish to} \\
&\qquad \text{consider over the planning horizon.}
\end{aligned}$$

We feel that it is important to recognize that $\overline{V}_t(R_t)$ is a forecast, using information available up through time t (just as we would from our demand forecast) of the value of resources in the future. Furthermore, we feel that $\overline{V}_t(R_t)$ is a sufficiently different type of forecast that it should be recognized as an information class distinct from Ω_t.

Finally, we feel that when an expert in operations expresses opinions about how a system should behave (represented using the pattern vector ρ_t), this is also a form of forecast that brings into play the experience of our expert. ρ_t can be thought of as a forecast of a decision that the model should make. It will not be a perfect forecast, as we will not in general be able to match these patterns perfectly, but just as we try to match historical patterns using a demand forecasting routine, these patterns are a forecast which reflects patterns of past decisions.

We have, now, four classes of information:

$$\begin{aligned}
K_t &= \text{the data knowledge at time } t, \\
\Omega_{[t,T^{ph}]} &= \text{a set of potential outcomes of exogenous information} \\
&\qquad \text{events in the future,} \\
\overline{V}_t(R_t) &= \text{forecasts of the impact of decisions now on the future,} \\
\rho_t &= \text{forecasts of future decisions, captured in the form} \\
&\qquad \text{of low-dimensional patterns.}
\end{aligned}$$

We can now formulate an information set I_t which will always include K_t, and can also include any combination of the remaining information sets. If we are solving the problem using only a rolling horizon procedure, we would specify $I_t = (K_t, \Omega_{[t,T^{ph}]})$. If we want to use dynamic programming, we would write $I_t = (K_t, \overline{V}_t(\Omega_t))$ to express the fact that our decision function depends on value functions which themselves depend on forecasts of future events. All four information classes would be written: $I_t = (K_t, \overline{V}_t(\Omega_t), \rho_t)$.

We see, then, that there are four fundamental information classes, and each brings into play specific algorithmic strategies. All the approaches assume that we are starting with a basic myopic model, and that we have algorithms for solving these problems. It is possible in theory to combine rolling horizon procedures and value functions, but value functions do offer a competing strategy for using forecasts of exogenous information processes to produce a forecast of a value of a resource in the future. Proximal point terms which incorporate expert knowledge can complement any model.

We can put all four information classes into a single decision function. Assume that we wish to manage equipment to move freight, including freight that is in the system now as well as forecasted freight within a planning horizon $\mathcal{T}^{ph}$. We could formulate a decision function that looks like

$$X_t^{\pi}(I_t) = \arg\max_{x_t \in \mathcal{X}_t} \sum_{t'=t}^{T^{ph}} \sum_{t''=t'}^{T^{ph}} \Bigg(c_{t',t''} x_{t',t''} + \sum_{t'' > T^{ph}} \overline{V}_{t,t''}\big(R^x_{t,t''}(x_t)\big) - H\big(\rho(x), \rho\big) \Bigg).$$

In this formulation, we are planning the flows of equipment within a planning horizon, using a point forecast of future information within this horizon. Equipment that becomes actionable past the end of the planning horizon is captured by a value function approximation, which we have represented as being separable over time periods. More commonly, optimizing over a planning horizon using a point forecast is done specifically to avoid the need for a value function. If value functions are used, these can be particularly nice mechanisms for handling stochastic forecasts. For this reason, we prefer to write our general purpose decision function as

$$X_t^{\pi}(R_t) = \arg\max_{x_t \in \mathcal{X}} \sum_{t'=t}^{T^{ph}} \Bigg(c_{t,t'} x_{t,t'} + \sum_{t' > T^{ph}} \overline{V}_{t,t'}\big(R^x_{t,t'}(x_t)\big) - H\big(\rho(x), \rho\big) \Bigg). \quad (61)$$

The idea of manipulating the information set available to a decision function is illustrated in Wu et al. (2003) in the context of the military airlift problem.

This problem, simply stated, involves assigning cargo aircraft to move "requests" which are loads of freight and passengers. The air mobility command currently uses a simulation package which captures a host of engineering details, but which uses very simple rules for choosing which aircraft should move each request. We use a version of these rules to represent the base policy. These rules work roughly as follows. The program maintains a list of aircraft in the order they are available, and requests to be moved, also in the order that they are available. Starting with the request at the top of the list, the program then determines whether the first available aircraft is eligible to move the request (there are a number of engineering restrictions). If not, it moves to the next aircraft on the list. As requests are satisfied, they are dropped from the list, and as aircraft are moved, they become available again in the future. Since a decision of which aircraft to choose does not compare different aircraft to find the best one, we can view this as a decision with very low information content. We refer to this as "rule-based, one aircraft, one request".

The next step up is to choose a request, and then find the best out of a list of aircraft using a cost function. The list of aircraft that are considered include only those which are known now and actionable now (that is, we are limited to aircraft in the vector R_{tt}). This strategy is a cost-based, myopic policy, using a single aircraft and a list of requests. Next is a strategy that considers a list of requests and a list of aircraft, which requires solving an assignment problem to determine the best assignment of each aircraft to each request. This strategy is also restricted to aircraft and requests that are known now, actionable now. The fourth information set is the same as the third, but now we are allowed to consider aircraft and requests that are known now but actionable in the future. The fifth information set includes value functions, which are used to capture the cost of putting an aircraft into an airbase where it might break down, incurring maintenance costs that can vary significantly between airbases.

The results of these simulations are shown in Figure 5, demonstrating steady improvement in costs as information is added to the decision function. We note that the value of increasing the information set depends on the problem. In addition, each increase in the information set adds to the computational requirements of the simulation. The use of value functions, for example, requires running the simulation iteratively, although it is often possible to estimate the value functions once and store these.

We did not include expert knowledge in this set of experiments since this information class exists only because the cost function is incomplete in some way. Imposing exogenous patterns through a proximal point term will always increase the costs. The appropriate measure here is the degree to which we are matching an exogenous pattern. To illustrate this process, we imposed a single exogenous pattern controlling the fraction of time that a particular aircraft type moved through a particular set of airbases. Without an exogenous pattern, this activity occurred approximately 20 percent of the time. Figure 6 shows the percentage of time this activity happened in the model, as a function of the weighting parameter θ. As θ is increased, the model produces results that are

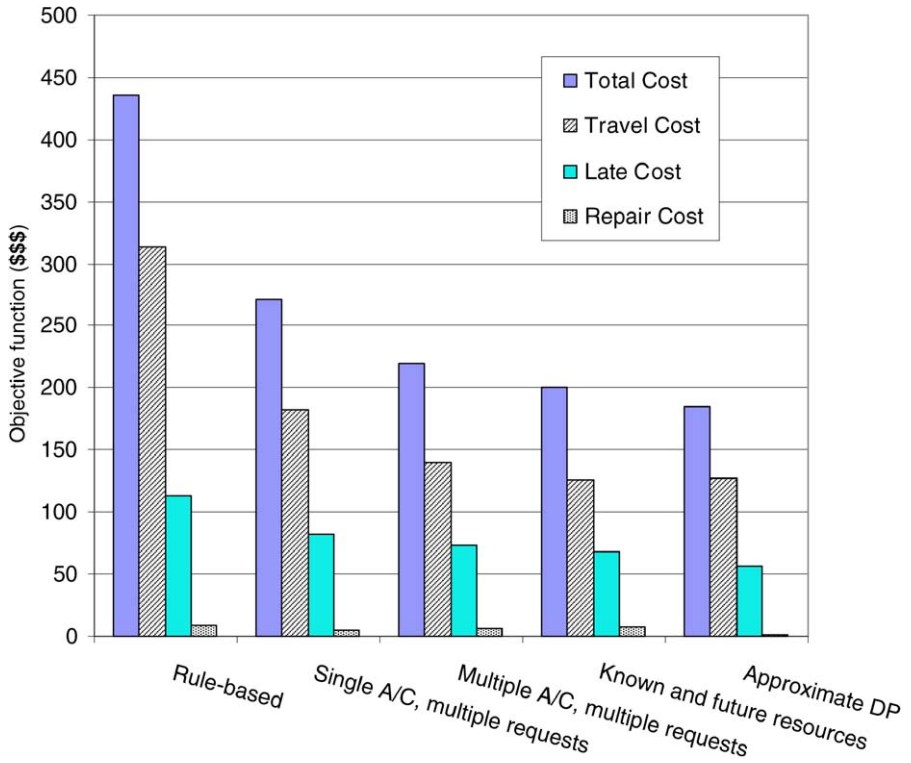


Fig. 5. Total costs, broken down between travel costs, late penalty costs, and repair costs, using four policies: (1) rule-based, chooses a single aircraft and single request and implements if feasible, (2) cost-based, considers a single aircraft and assigns to the best of a list of requests based on cost, (3) minimizes the costs of assigning a set of aircraft to a set of requests, limited to those that are both known and actionable now, (4) same as (3), but now includes aircraft and requests that are known now but actionable in the future, and (5) includes value functions to guide the choice of aircraft (from Wu et al., 2003).

closer to the desired pattern, although there is a limit to this. In practice, if an expert says that we can do something 10 percent of the time, it is very likely that a solution that does it 15 or 20 percent of the time would be acceptable.

We are not trying to make the argument that all four information classes must be used in all models, but the exclusion of any information class should be recognized as such, and our experience has been that each will add value, especially when working on complex problems.

8.6 Bibliographic notes

There is a vast literature on myopic models and rolling horizon procedures that use point forecasts (virtually any deterministic dynamic model falls in this class). A number of authors have developed algorithms for exploiting the dy-

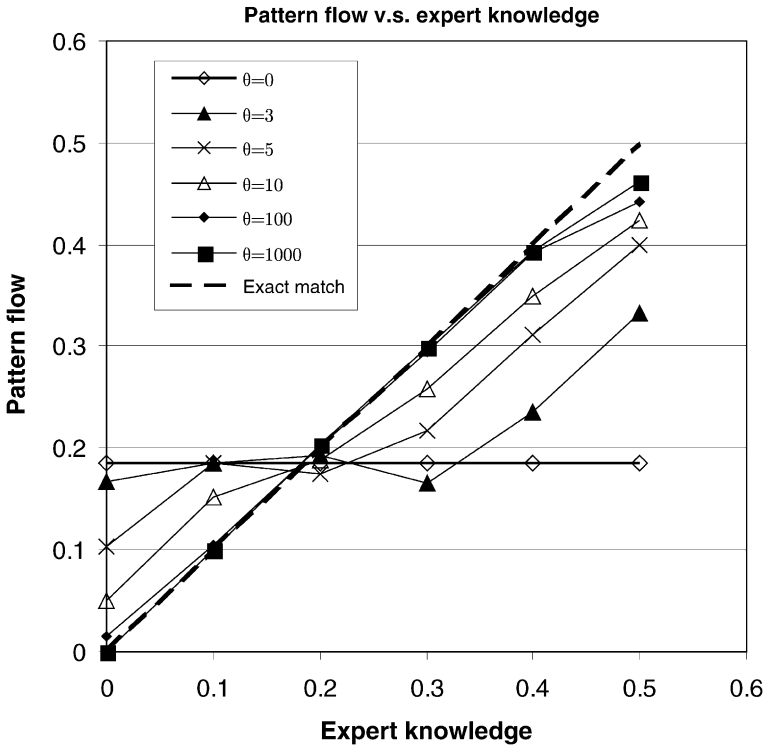


Fig. 6. Actual versus desired flow of a particular aircraft through a particular airbase, as a function of the pattern weight parameter θ.

namic structure of these models (see, for example, Aronson and Chen, 1986; Aronson and Thompson, 1984) although this work has been largely replaced by modern linear programming techniques. Stochastic programming models and algorithms are reviewed in Infanger (1994), Kall and Wallace (1984), Birge and Louveaux (1997), Sen and Higle (1999). There are numerous texts on dynamic programming. Puterman (1994) is the best overview of classical methods of Markov decision processes, but covers only discrete dynamic programs, with algorithms that work only for very low dimensional problems. The use of the post-decision state variable has received very little attention. It is used implicitly in an exercise in Bellman's original text (Bellman, 1957), but not in a way that identifies its importance as a computational device. Godfrey and Powell (2002) use the post-decision state variable explicitly in the development of an approximate dynamic programming algorithm for fleet management, extending an approximation method used in a series of papers (Powell, 1987; Cheung and Powell, 1996) where the technique was used implicitly. Van Roy (2001) introduces pre- and post-decision state variables, taking advantage of the fact that the post-decision state variable avoids the expectation in the determination of an optimal action. This property is not mentioned in any of the

standard textbooks on exact or approximate dynamic programming (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998), but it appears to be a powerful technique for high-dimensional resource allocation problems. The explicit introduction of four classes of information appears to have been first done in Powell et al. (2001). The investigation of low-dimensional patterns is investigated in Marar et al. (2006) and Marar and Powell (2004), and is applied to the military airlift problem in Powell et al. (2004).

9 Approximating value functions in dynamic programming

Section 8.3 introduces the idea of using approximate value functions, but did not fill in any of the details of how to do this. Approximate dynamic programming is an emerging field that must be proven experimentally on different problem classes, just as different heuristics have to be tested and proven on different problems. This section, which is easily skipped on a first reading, provides an introduction to the current state of the art in approximate dynamic programming, focused on the types of problems that arise in transportation and logistics.

Section 9.1 focuses on a problem where there is a single resource which might be characterized by a vector of attributes. This treatment lays the foundation for solving problems with many resources. Section 9.2 describes strategies for estimating continuous value function approximations, which are a central strategy for avoiding "curse of dimensionality" problems in dynamic programs.

9.1 Discrete value function approximations for a single resource

Fundamental to the challenge of allocating multiple resources is the problem of allocating a single resource. This helps us illustrate some basic strategies, but also highlights some issues that arise in more complex problems.

Our discussion is helped by a simple example. Assume we have the case of a nomadic trucker who moves around the country from state to state (within the United States). His attribute vector (for the moment) is the "state" that he is in (such as New York). From a state, he has to choose from a number of loads (requests to move freight) that he learns about after his arrival. The load he chooses should reflect the profit he expects from the load, plus the value of being in the state that the load terminates in (this will maximize our trucker's long term profits). If we wanted to make our problem more realistic, we would include attributes such as the driver's home domicile, how many days he has been away from his home, and the maintenance status of his truck (was there an equipment failure during his last trip that would require a repair at the destination?).

We let $a_{t-1} \in \mathcal{A}$ be the attribute of our driver at time $t-1$ and $V_{t-1}(a_{t-1})$ be the expected future value of a driver that has attribute a at time $t-1$.

We can formulate this problem in two ways: as we have above using R_t as the state variable and x_t as our decision, or using a_t as the state variable and d as the decision where the set of potential decisions $\mathcal{D}_t$ depends on the information available at time t. Recalling that $a^{\mathrm{M}}(t, a_{t-1}, d)$ is the attribute vector produced by applying the decision d to a resource with attribute a_{t-1}, we can write Bellman's equations as

$$V_{t-1}(a_{t-1}) = E\Big\{\max_{d\in\mathcal{D}_t}\big(c_{t,a_{t-1},d} + V_t\big(a^{\mathrm{M}}(t, a_{t-1}, d)\big)\big)\big|a_{t-1}\Big\}. \tag{62}$$

Equation (62) is the more natural way to formulate this specific problem, but it disguises the relationship to problems with multiple resources. A natural way to bring the two formulations together is as follows. Since $\sum_{a\in\mathcal{A}} R_{ta} = 1$, we can write $V_t(R_t) = V_t(a_t) = \sum_{a\in\mathcal{A}} R_{ta}V_t(a)$, where $R_{ta_t} = 1$. Thus, for the single resource problem, $V_t(R_t)$ is linear in R_t.

If the attribute space $\mathcal{A}$ is not too large, we can solve Equation (62) using classical backward dynamic programming techniques. Not surprisingly, these problems do not generally exist in practice (the problems with resources with simple attribute vectors, such as containers or freight cars, are exactly the problems with many resources). The more interesting problems involve managing people (which tend to exhibit complex attribute vectors) or in multistop vehicle routing problems, where the attribute vector has to capture the stops that a vehicle has already made. For these problems, we resort to our approximation strategy, where it is fairly obvious that a linear approximation is the appropriate form. This gives us

$$X_t^{\pi}\big(R_{t-1}^x, \omega^n\big) = \max_{x_t\in\mathcal{X}_t(\omega^n)}\left(c_t x_t + \sum_{a'\in\mathcal{A}} \bar{v}_{ta'}^n R_{ta'}^x(x_t)\right), \tag{63}$$

where $R_{ta'}^x(x_t) = \sum_{a\in\mathcal{A}}\sum_{d\in\mathcal{D}_a} x_{t,a,d}\delta_{a'}(t, a, d)$ is our post-decision resource vector. Our decision function X_t^{π} is written as depending on R_{t-1}^x and ω, but it actually depends on $R_t = R_{t-1}^x + \widehat{R}_t(\omega)$.

The result of solving (63) is a random estimate $\hat{v}_{ta}^n$ of the value of a resource with attribute a_{t-1} with the information available at time t. If we were using the simple one-pass algorithm, we could now update $\bar{v}_{t-1}^{n-1}$ using

$$\bar{v}_{t-1,a}^n = \big(1-\alpha^n\big)\bar{v}_{t-1,a}^{n-1} + \alpha^n\hat{v}_{ta}^n. \tag{64}$$

Again we note that $\hat{v}_{ta}^n$ is indexed by t because it contains information from time t, while $\bar{v}_{t-1,a}^n$ represents an approximation of an expectation over all possible outcomes of W_t, and therefore implicitly contains only information up to time $t-1$.

If we were to use the two-pass version of the algorithm, we would simulate the trajectory of our nomadic trucker using the policy $(X_{t'}^{\pi})_{t'\geqslant t}$. We would then compute $\hat{v}_{ta}^n$ as the cost of the trajectory over the interval (t, T), and then update $\bar{v}$ as we did in Equation (64).

These procedures are known in the approximate dynamic programming literature as temporal-differencing methods (Sutton, 1988; Sutton and Barto, 1998; Bertsekas and Tsitsiklis, 1996) where they are normally described in the context of steady state problems. They can also be viewed as examples of stochastic gradient algorithms. To see this, we can write the problem of estimating the exact function $V_t(a)$ as one of solving

$$\min_{v_{ta}} f(v) = E\left\{\frac{1}{2}(v_{ta} - \widehat{V}_{ta})^2\right\}, \tag{65}$$

where $\widehat{V}_{ta}$ is a random variable representing a measurement of the function with noise. We note that $\widehat{V}_{ta}$ is an unbiased estimate only when we use the two-pass version of the algorithm. Equation (65) can be solved by dropping the expectation, taking a Monte Carlo sample and then applying a standard gradient algorithm

$$\begin{aligned} v_{ta}^n &= v_{ta}^{n-1} - \alpha^n \nabla_v f\left(v_{ta}^{n-1}, \omega^n\right) \\ &= v_{ta}^{n-1} - \alpha^n \left(v_{ta}^{n-1} - \widehat{V}_{ta}^n(\omega)\right) \\ &= \left(1 - \alpha^n\right) v_{ta}^{n-1} + \alpha^n \widehat{V}_{ta}^n(\omega). \end{aligned} \tag{66}$$

The procedure (66) is known as a stochastic approximation algorithm (Robbins and Monro, 1951; see Kushner and Yin, 1997, for a modern treatment of the field). In addition to some technical requirements on the properties of the function being estimated, the proof requires that $\widehat{V}_{ta}^n(\omega)$ be an unbiased estimate of V_{ta} in the limit (biases are allowed initially as long as they go to zero reasonably quickly), and two conditions on the stepsizes: $\sum_{n=1}^{\infty} \alpha^n = \infty$ and $\sum_{n=1}^{\infty} (\alpha^n)^2 < \infty$. These requirements basically require that the stepsize decrease to zero arithmetically, as in $\alpha^n = 1/n$.

This simple problem highlights an issue that plagues all approximate dynamic programming methods. Assume that we start with an initial approximation $\bar{v}_{ta} = 0$ for all t and a. In the case of our nomadic trucker, this means that initially, he will prefer long loads with high initial rewards over shorter loads with low initial reward. In fact, a shorter load might be more profitable in the long run if it takes him to a region with high profit loads. If we only update the values of states that he visits, then he will tend to only visit states that he has already visited before.

One way to avoid this behavior is to begin by initializing $\bar{v}_t$ with an upper bound. This way, if our trucker turns down a load to a state that he has never visited, then we can be assured that we are not turning away an option that might prove to be quite good. In practice, this strategy works poorly, because it is more likely that he will always choose the load that goes to a state that he has never visited (for our problem, it is not just a state, but also a state at a point in time). If the attribute vector includes other dimensions, then our attribute space $\mathcal{A}$ can grow very quickly (a nice example of the curse of dimensionality).

The problem of visiting states in order to estimate the value of being in the state is called the *exploration* problem in dynamic programming. There is no

cure for this problem, but there are strategies that help overcome it. One is to use "off-policy iterations" which means to choose actions that are not optimal given the current approximations, but forces the system to visit new states. In real problems, the number of states and actions tends to make such strategies impractical (at a minimum, they have very slow convergence). A more effective strategy is to exploit problem structure. For example, assume that one attribute of our truck driver is the number of days away he has been from home, and suppose that we are able to establish that the value of the driver declines monotonically with this quantity. We can then use this structural property to help estimate the values of states we might never have visited.

The problem of estimating high-dimensional functions has open theoretical and practical questions. From a theoretical perspective, existing proofs of convergence for discrete value functions requires visiting all states infinitely often. This generally requires an algorithm that combines optimizing iterations (where we choose an action based on our approximate policy) with learning iterations (where we might choose actions at random). From a practical perspective, learning steps are only of value with low-dimensional action spaces. Practical considerations tend to focus on the rate of convergence, which tend to be highly problem dependent.

9.2 Continuous value function approximations

A popular strategy for dealing with the curse of dimensionality has been to replace the value function with a continuous approximation (Bellman and Dreyfus, 1959). Particularly appealing are the use of low-order polynomials. In this section, we describe several strategies for estimating continuous approximations.

9.2.1 Basis functions

A strategy for estimating a continuous approximation of a value function is to use the concept of basis functions. Assume, for example, that we are managing a fleet of containers of different types, and we believe that the number of containers of each type at each location is an important determinant of the value of a particular allocation. We could devise a function, call it $\phi_a(R) = \sqrt{R_{ta}}$. $\phi_a(R)$ takes the square root of this quantity because we believe that there are diminishing returns from additional capacity. In this case, we would call the square root of the number of containers with attribute a "a feature" of our problem. Another feature might be the total number of containers of all types at a location i. We might then create a new class of functions $\phi_i(R) = \sqrt{\sum_{a \in \mathcal{A}_i} R_{ta}}$, where $\mathcal{A}_i$ is the set of attribute vectors that correspond to resources at a particular location i. Another feature could be the number of containers minus the forecasted number of orders. It is up to our imagination and insight to identify what appear to be important features. Let $\mathcal{F}$ be the set of all these functions, known as *basis functions* or *features*. We can formulate

an approximate value function as a linear combination of our basis functions

$$\overline{V}(s|\theta) = \sum_{f\in\mathcal{F}} \theta_f \phi_f(s). \tag{67}$$

Now the challenge is to determine the appropriate weights $(\theta_f)_{f\in\mathcal{F}}$. We illustrate two strategies for doing this. The first is to simply apply a stochastic gradient algorithm. We generalize slightly the minimization problem in (65) as follows

$$\min_{\theta} E\left\{\frac{1}{2}\big(\overline{V}^{n-1}(s|\theta) - \widehat{V}(s)\big)^2\right\}.$$

Solving this using a stochastic gradient algorithm produces updates of the form

$$\theta^n = \theta^{n-1} - \alpha^n\big(\overline{V}^{n-1}(s|\theta) - \widehat{V}(s,\omega)\big)\nabla_\theta \overline{V}(s,\omega|\theta). \tag{68}$$

The challenge that this class of update procedures faces is the problem of scaling the stepsize, since the units of θ and the units of the updating term will be completely different. An alternative strategy that avoids this problem takes a sample $\widehat{\Omega}$ and then finds θ that minimizes the deviation between the approximation and the sample

$$\theta^n = \arg\min_{\theta} \frac{1}{|\widehat{\Omega}|}\sum_{\omega\in\widehat{\Omega}}\left(\frac{1}{2}\big(\overline{V}^{n-1}(s,\omega|\theta) - \widehat{V}(s,\omega)\big)^2\right). \tag{69}$$

We can then use the value of θ that solves (69), or smooth this estimate with the previous estimate.

An introduction to the use of basis functions can be found in Bertsekas and Tsitsiklis (1996), with an in-depth treatment of the theoretical foundation in Tsitsiklis and Van Roy (1997). However, there is virtually no experimental work in the area of transportation and logistics. There are two practical challenges: identifying effective basis functions that capture the important properties of the problem, and the second is ensuring that the rate of convergence is sufficiently fast.

9.2.2 Auxiliary functions

Consider the case of a continuous resource allocation problem (where R_t is continuous). This means that we can find a stochastic gradient of $\widetilde{V}_t^n(R_t(\omega^n))$ in Equation (49). Let

$$\hat{v}_t^n = \nabla_{R_t}\widetilde{V}_t^n\big(R_t\big(\omega^n\big)\big)$$

be a stochastic gradient of the function. Now let $\overline{V}_t^0(R_t)$ be a conveniently chosen continuous approximation (perhaps a low-order polynomial). We can improve our approximation using the updating equation

$$\overline{V}_t^n(R_t) = \overline{V}_t^{n-1}(R_t) + \alpha^n\big(\hat{v}_t^n - \nabla_{R_t}\overline{V}_t^{n-1}(R_t^n)\big)R_t. \tag{70}$$

This is the primary step of the SHAPE algorithm (Cheung and Powell, 2000). It starts with an arbitrary (but presumably carefully chosen) initial approximation, and then tilts it using stochastic gradient information. The second term of Equation (70) is a linear updating term that adds the difference between the stochastic gradient of the real function and the exact gradient of the approximate function. Each updated function has the same basic form as the initial approximation. This can be a nice simplification if it is necessary to precompute derivatives.

The SHAPE algorithm is optimal for two-stage, continuously differentiable problems. For nondifferentiable problems, and for multistage problems, it is an approximation. In principle the initial approximation could be estimated from a set of basis functions, allowing both methods to be used as a hybrid strategy.

Computational testing of these methods is limited. Their success appears to be highly dependent on a good choice of initial approximation. Simply choosing low-order polynomials because they are convenient is not likely to be successful.

9.2.3 Linear functional approximations

We have already seen in the case of managing a single resource that the value function can be written as a linear function of the resource vector. Linear approximations can also be effective when we are managing very large numbers of resources. For example, shipping companies might manage fleets of hundreds of thousands of containers. Our ability to effectively manage resources depends primarily on our ability to estimate the slope of the function, rather than the function itself. When there are large numbers involved, the slopes tend to be fairly stable.

Throughout this section, we assume that $\hat{v}_{ta}$ is a stochastic gradient of either $\widetilde{V}_t(R_t)$ with respect to R_{ta} (Equation (49)), or of $V_t^{\pi}(R_t, \omega)$ (Equation (52), computed using a backward pass). The choice of which version to use is primarily one of trading off speed of convergence with algorithmic complexity (backward passes can be hard to implement but can dramatically improve results). Given an estimate of the slope, we would then smooth on the slope (as in Equation (64)) to obtain an estimate of the slope, producing the following approximation

$$\widetilde{V}_t^n\big(R_t^n(\omega^n)\big) = \max_{x_t \in \mathcal{X}_t(\omega^n)} \left(c_t x_t + \sum_{a \in \mathcal{A}} \bar{v}_{ta}^{n-1} R_{t,a}^x(x_t) \right). \tag{71}$$

Using our $\delta(\cdot)$ function, we can write Equation (71) as

$$\begin{aligned} &\widetilde{V}_t^n\big(R_t^n(\omega^n)\big) \\ &\quad = \max_{x_t \in \mathcal{X}_t(\omega^n)} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} c_{tad} x_{tad} \\ &\qquad + \sum_{a' \in \mathcal{A}} \bar{v}_{ta'}^{n-1} \left(\sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} \delta_{t,a'}(t, a, d) x_{tad} \right) \end{aligned}$$

$$= \max_{x_t \in \mathcal{X}_t(\omega^n)} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} c_{tad} x_{tad} + \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} \bar{v}^{n-1}_{t,a^{\mathrm{M}}(t,a,d)} x_{tad} \tag{72}$$

$$= \max_{x_t \in \mathcal{X}_t(\omega^n)} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} c_{tad} x_{tad} + \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} \bar{v}^{n-1}_{t,a^{\mathrm{M}}(t,a,d)} x_{tad} \tag{73}$$

$$= \max_{x_t \in \mathcal{X}_t(\omega^n)} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} \big(c_{tad} + \bar{v}^{n-1}_{t,a^{\mathrm{M}}(t,a,d)}\big) x_{tad}. \tag{74}$$

We obtain Equation (72) by using $\delta_{t,a'}(t, a, d) = 1$ for $a' = a^{\mathrm{M}}(t, a, d)$ and 0 otherwise, producing the more compact expression in (73). Equation (74) is obtained by simply rearranging terms, which shows that we are now solving a problem with the same structural form as the original myopic problem, with a modified set of costs.

This structure is especially important in transportation problems since the original myopic problem might be fairly difficult. If we are lucky, it is an assignment problem, transportation problem or some linear program. But in some settings it is a set partitioning problem, vehicle routing problem, or integer multicommodity flow problem. For these problems, it is especially helpful when the value function approximation does not complicate the problem further.

Not surprisingly, linear value functions do not always work well. It is very hard to tune a linear function to attract just the right number of resources. Nonlinear functions are simply more stable, but as we show below, they introduce considerable complexity.

9.2.4 Separable piecewise-linear functional approximations

When we are managing discrete resources (freight cars, trailers, people, aircraft, locomotives) it is especially important to obtain integer solutions. This is simplified considerably when we use separable, piecewise-linear approximations which are linear for noninteger values of the number of resources. For this strategy, we write our approximation as

$$\overline{V}_t(R_t) = \sum_{a \in \mathcal{A}} \overline{V}_{ta}(R_{ta}),$$

where

$$\overline{V}_{ta}(R_{ta}) = \sum_{r=0}^{\lfloor R_{ta} \rfloor - 1} \bar{v}_{ta}(r) + \bar{v}_{ta}\big(\lfloor R_{ta} \rfloor\big)\big(R_{ta} - \lfloor R_{ta} \rfloor\big). \tag{75}$$

When we introduce this approximation in Equation (49) we obtain

$$\widehat{V}^n_t\big(R^n_t(\omega^n)\big) = \max_{x_t \in \mathcal{X}_t(\omega^n)} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} c_{tad} x_{tad} + \sum_{a' \in \mathcal{A}} \overline{V}_{ta'}(R_{ta'}) \Bigg(\sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}_a} \delta_{t,a'}(t, a, d) x_{tad} \Bigg). \tag{76}$$

Assume that the original myopic problem is a pure network. We are interested in the conditions under which the introduction of value functions destroys this network structure. The constraint set $\mathcal{X}_t(\omega^n)$ typically contains upper bounds representing vehicle capacity constraints, network constraints, or simply constraints on the number of tasks to be served. Let u_{td} be the upper bound on the total number of resources acted on by a particular decision type

$$\sum_{a \in \mathcal{A}} x_{tad} \leqslant u_{td}. \tag{77}$$

Equation (77) has the effect of creating a bundle constraint. This does not pose a problem when we use linear value functions, as we demonstrated that any linear value function is equivalent to solving the original problem with a modified set of costs. But when we use nonlinear value functions, the network structure is destroyed.

Of particular interest are problems that satisfy the Markov property:

Definition 9.1. A resource allocation problem satisfies the Markov property if

$$a^{\mathrm{M}}(t, a, d) = a^{\mathrm{M}}(t, a', d), \quad \forall a' \in \mathcal{A}.$$

The Markov property means that the attributes of a resource after being acted on are independent of the attributes of the resource. A special class of problems that satisfy the Markov property are known as single commodity flow problems. In this problem class, the attribute vector consists purely of the location of a resource. A task consists of an origin and a destination, so after completing the task, the attribute of the resource equals the destination of the task (and is therefore independent of the origin). Another example arises when we have to maintain an attribute of whether a container is clean or dirty. Assume there is freight that is classified as clean or dirty. Only a clean container can move clean freight, but any container can move dirty freight. However, if a clean car moves dirty freight, then it becomes dirty. Such a problem would also satisfy the Markov property.

When a problem possesses the Markov property, the attributes of the resource after being acted on are completely determined by the decision rather than the attributes of the resource. As a result, we can create a node for resources that have been acted on by a particular type of decision, and we do not have to track the identity of the origin of the resource after it completes a task (or any other action). As a result, it is possible to formulate the problem as a pure network.

When problems do not possess the Markov property, the resulting problem will not, in general, be a pure network. But the use of piecewise linear, separable value function approximations appears to produce integer multicommodity flow problems with very tight LP relaxations. Experimental evidence is that when these problems are solved as continuous linear programs, the solution returned is integer the vast majority of the time (99.9 percent).

9.3 Bibliographic notes

The field of approximate dynamic programming is relatively young. The first significant reference is Bertsekas and Tsitsiklis (1996) which gives a general introduction to a variety of methods, but without presenting specific algorithms. Sutton and Barto (1998) approach the topic from the perspective of reinforcement learning. Temporal-difference learning, a form of stochastic approximation procedure, was first introduced by Sutton (1988). The vast majority of the literature on dynamic programming appears to focus on discrete representations of the value function, but there has been a steady literature on the use of continuous approximations, beginning as early as Bellman and Dreyfus (1959). Tsitsiklis and Van Roy (1997) investigate in depth the convergence properties of temporal-difference methods using continuous approximations. The SHAPE algorithm was introduced by Cheung and Powell (2000). The theory behind the estimation of linear value function approximations is based on the seminal work on stochastic approximation methods by Robbins and Monro (1951) and Blum (1954) (see Kushner and Yin, 1997, for a modern review of these techniques). The use of separable, piecewise linear approximations was first introduced by Cheung and Powell (1996) who used a static approximation. Godfrey and Powell (2001) proposes an adaptive learning algorithm for piecewise linear, separable approximations in the context of two-stage resource allocation problems, which is extended in Godfrey and Powell (2002) for multistage, single commodity problems. Topaloglu and Powell (2006) shows that these techniques also work quite well for multicommodity flow problems with a range of forms of substitution.

10 The organization of information and decisions

Another important issue in the modeling of decisions and information is how they are organized. Large transportation operations are typically characterized by different decision makers controlling different parts of the problem with different information. In this section, we provide some basic notation to model how organizations are controlled.

We adopt the convention that a decision-maker is referred to as an "agent" which is assumed to have sole control over a set of decisions. We let

$$\mathcal{Q} = \text{set of decision-making agents},$$

$$\mathcal{D}_q = \text{the set of decisions controlled by agent } q.$$

Our first challenge is to formalize the decision sets $\mathcal{D}_q$. We assume that the sets $(\mathcal{D}_q)_{q\in\mathcal{Q}}$ are mutually exclusive and collectively exhaustive.

There are three dimensions to the set $\mathcal{D}_q$: the types of decisions an agent can make, the types of resources an agent can act on, and the time periods at which the decisions might be implemented. For example, a locomotive planner for a railroad might have the responsibility for planning the movements

of locomotives on Monday, Tuesday, and Wednesday. He can only control locomotives that are in service; a local maintenance supervisor has control over locomotives that are classified as being "in maintenance" (it is common for a locomotive planner to call the local maintenance supervisor to determine when a locomotive that is in the shop will become available). Furthermore, he can make decisions about coupling and uncoupling locomotives to and from trains, but does not make decisions about buying, selling, leasing, and storing locomotives. To capture this process, we let

$$\mathcal{A}_q = \text{subset of the attribute space for subproblem } q, \text{ where}$$
$$\bigcup_{q \in \mathcal{Q}} \mathcal{A}_q = \mathcal{A} \text{ and } \mathcal{A}_{q_1} \cap \mathcal{A}_{q_2} = \phi \text{ when } q_1 \neq q_2.$$

Implicit in the definition of the attribute space is:

$\mathcal{C}_q^{\mathrm{D}}$ = set of control classes associated with subproblem q. As a rule, a subproblem is formulated for a specific type of decision, so the set $\mathcal{C}_q^{\mathrm{D}}$ is implicit in the formulation of the problem,

$\mathcal{D}_a^c$ = set of decisions in control class c that can be applied to a resource with attribute vector a,

$\mathcal{T}_q^{ih}$ = the *implementation horizon* for subproblem q. This is the set of time periods during which subproblem q controls the decisions. Since time is a dimension of the attribute vector, we may state that $a \in \mathcal{A}_q \Rightarrow a_{\text{actionable}} \in \mathcal{T}_q^{ih}$.

It is important to emphasize that the subsets $\mathcal{A}_q$ do not necessarily (or even typically) imply a geographical decomposition, although this certainly can be the case. At a railroad, locomotives are managed regionally (with some network-wide coordination); freight cars are managed network-wide, but are decomposed by freight car type.

Our decision set for agent q is now given by

$$\mathcal{D}_q = \text{subset of decisions in subproblem } q$$
$$= \left\{ d \in \mathcal{D}_a^c, c \in \mathcal{C}_q^{\mathrm{D}}, a \in \mathcal{A}_q, a_{\text{actionable}} \in \mathcal{T}_q^{ih} \right\}.$$

The definition of $\mathcal{D}_q$ includes the time periods for which the decisions may be implemented, which produces a more compact notation. In some cases, the indexing of when a decision is to be implemented needs to be made explicit, so we can use (t, q) to capture the combined information.

Before, the information available to make a decision was represented implicitly either by indexing time t, or explicitly through a state variable such as S_t or R_t. When we model multiagent systems, we have to be more explicit. For

this reason, we define:

$$I_{tq} = \text{the information content of subproblem } q \text{ at time } t.$$

It is best to think of I_{tq} as a set of functions that return data that is needed to make a decision: the amount of equipment, people, and freight, available now and in the future (recall that the index t here refers to information content, not actionability).

Another dimension of multiagent systems is capturing the property that decisions by agent q can have an impact on another agent q' (for example, sending a piece of equipment from one region to the next, or when the person who buys and sells equipment changes the amount of equipment that a planner can work with). We capture these interactions by defining the following:

Definition 10.1. The *forward reachable set* $\vec{\mathcal{I}}_q$ of subproblem q is the set of subproblems q' with resource states $a' \in \mathcal{A}_{q'}$ that can be reached by implementing a single decision $d \in \mathcal{D}$ on at least one state $a \in \mathcal{A}_q$. More precisely, the forward reachable set of subproblem q is

$$\vec{\mathcal{I}}_q = \big\{q' \in \mathcal{Q} \setminus q \mid \exists a \in \mathcal{A}_q, d \in \mathcal{D}_q, \text{ where } M(a, d, \cdot) \mapsto (a', \cdot, \cdot), a' \in \mathcal{A}_{q'}\big\}. \tag{78}$$

For completeness, we also define:

Definition 10.2. The *backward reachable set* $\overleftarrow{\mathcal{I}}_q$ of subproblem q is the set of all subproblems for which subproblem q is forward reachable. More precisely, the backward reachable set of subproblem q is

$$\overleftarrow{\mathcal{I}}_q = \{q' \in \mathcal{Q} \setminus q \mid q \in \vec{\mathcal{I}}_{q'}\}. \tag{79}$$

This allows us to define decision vectors for each agent:

$$\begin{aligned}
x_{taa'} &= \sum_{d \in \mathcal{D}_q} x_{tad} \delta_{a'}(t, a, d), \\
x_{tqa'} &= \{x_{taa'}, a \in \mathcal{A}_q\}, \\
x_{tqq'} &= \{x_{tqa'}, a' \in \mathcal{A}_{q'}\}, \\
x_{tq} &= \{x_{tqq'}, q' \in \vec{\mathcal{I}}_q\}.
\end{aligned}$$

We adapt our notation for making decisions by defining

$$X^{\pi}_{tq}(I_{tq}) = \text{a function that determines } x_{tad} \text{ for } a \in \mathcal{A}_q, d \in \mathcal{D}_q \text{ using the information available in } I_{tq},$$

where I_{tq} is the information known by agent q at time t. As we did before, we can create a general purpose decision function that uses all four information

classes

$$X^{\pi}_{tq}(I_{tq}) = \arg\max_{x_{tq}} c_{tq}x_{tq} + \sum_{q' \in \overline{\mathcal{I}}_q} \overline{V}_{tqq'}\big(R_{tq'}(x_{tqq'})\big) - H\big(\rho(x_{tq}), \rho_q\big). \tag{80}$$

Shapiro and Powell (2006) describe this strategy in depth and give a method for estimating the value functions $\overline{V}_{tqq'}(R_{tq'}(x_{tqq'}))$.

11 Illustrative models

We illustrate these ideas in the context of three problem areas: rail car distribution, load matching for truckload motor carriers, and batch processes that arise in less-than-truckload motor carriers. The car distribution problem is characterized by a low-dimensional attribute space (it is roughly a multicommodity flow problem) with an interesting set of dynamic information processes. The load matching problem introduces the challenge of modeling people (characterized by high-dimensional attribute spaces). Also, we demonstrate the modeling of a two-layer resource allocation problem (drivers and loads). Despite their differences, both of these problems have the flavor of time-staged linear programs. The batch processes that arise in less-than-truckload motor carrier illustrates how we can handle the types of fixed-charge problems that arise in network design.

11.1 Dynamic flow problems for rail car distribution

Our car distribution problem provides an opportunity to illustrate the notation that we have introduced. Section 11.1.1 provides a model of an actual car distribution system developed for a major railroad. Then, Section 11.1.2 illustrates several optimization models that are used in practice. Section 11.1.3 summarizes the results of some experiments run with an actual dataset from a major railroad. Section 11.1.4 closes with some observations about this application.

11.1.1 A dynamic model

The car distribution problem basically involves assigning cars to orders. Cars can be moved empty or assigned to orders and then moved loaded. Cars or orders that are not acted on are held. Orders that are not satisfied within a reasonable time (typically within the same week) are considered lost, but in practice some orders might simply be held to the next week. When a car becomes empty at a shipper, the railroad has roughly three options which are illustrated in Figure 7: assign the car to a pending order, move the car to a "regional depot" from which it can be used to satisfy orders from other shippers

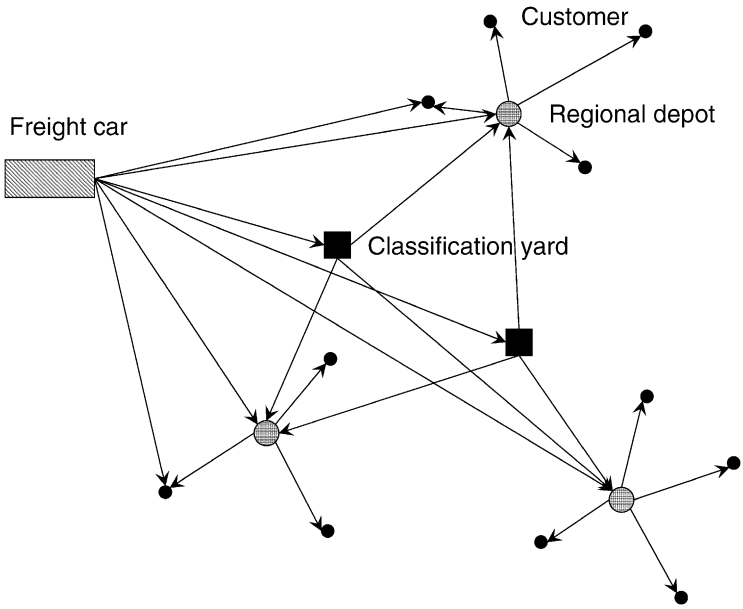


Fig. 7. Car distribution through classification yards.

in the region, or send the car to a "classification yard". A classification yard is a major sorting facility, from which cars might be sent to anywhere in the network.

Our problem, then, involves two "resource classes", namely cars and orders. We model cars using

a = the vector of attributes characterizing a car,

$\mathcal{A}$ = the set of attributes of the cars,

$R^{\mathrm{c}}_{t,t'a}$ = the number of cars with attribute a that we know about at time t that will be available at time t'. The attribute vector includes the location of the car (at time t') as well as its characteristics,

$$R^{\mathrm{c}}_{t,t'} = \left(R^{\mathrm{c}}_{t,t'a}\right)_{a\in\mathcal{A}},$$

$$R^{\mathrm{c}}_{t} = \left(R^{\mathrm{c}}_{t,t'}\right)_{t'\in\mathcal{T}}.$$

For our problem, we model the information process (the arrival of information and the making of decisions) in 24-hour increments. By contrast, we model the physical movement of resources in continuous time. As a result, our attribute

vector consists of

$$a = \begin{bmatrix} a_{\text{location}} \\ a_{\text{car\_type}} \\ a_{\text{actionable}} \end{bmatrix}.$$

The element of $a_{\text{actionable}}$ gives the time, down to the minute, of when a car is able to move. Note that there are some important modeling decisions in how the attribute is handled. Assume, for example, that a car at origin o has to pass through classification yards (where cars are sorted and can be rerouted) at i and j on its path to destination d. If the car is sent to d with expected arrival time t', we can update its attribute vector to include $a_{\text{location}} = d$ and $a_{\text{actionable}} = t'$. This means that we will not consider rerouting the car at i or j. On the other hand, we might wish to consider rerouting, but we do not want to ignore that it is already on a path to d. The railroad generates a piece of paper called a work order that routes the car to d; if we reroute the car to a different destination, this paperwork has to be changed. If we wish to allow rerouting, but without forgetting the original destination of a car, we would simply add a new dimension to a which we might call $a_{\text{destination}}$ which would be the ultimate destination of the car, and we would interpret a_{location} as the next location where a car is able to be rerouted (so, as the car departs from o heading to i, we would set $a_{\text{location}} = i$ and $a_{\text{destination}} = d$).

Orders are modeled in a similar way using

b = the vector of attributes characterizing an order,

$\mathcal{B}$ = the set of attributes of an order, including the number of days into the future on which the order should be served (in our vocabulary, its actionable time),

$R^{\text{o}}_{t,t'b}$ = the vector of car orders with attribute $b \in \mathcal{B}$ that we know about at time t which are needed at time t'. $R^{\text{o}}_{0,bt'}$ is the set of orders that we know about now,

$$R^{\text{o}}_{t,t'} = \left(R^{\text{o}}_{t,t'b}\right)_{b\in\mathcal{B}},$$

$$R^{\text{o}}_{t} = \left(R^{\text{o}}_{t,t'}\right)_{t'\in\mathcal{T}}.$$

The attributes of orders consist of

$$b = \begin{bmatrix} b_{\text{origin}} \\ b_{\text{destination}} \\ b_{\text{car\_type}} \\ b_{\text{actionable}} \end{bmatrix}.$$

Here $b_{\text{car\_type}}$ represents the preferred car type, but substitution opportunities exist. $b_{\text{actionable}}$ is the exact time that the order is available to be picked up. We note that when an order becomes known, it is not necessarily the case that all the attributes of an order become known at the same time. So we may

adopt a convention that, say, $b_{\text{destination}} =$ '-' means that the destination is not yet known.

Our resource, vector, can now be written

$$R_t = \left(R_t^{\text{c}}, R_t^{\text{o}}\right).$$

One of the challenges of car distribution is dealing with a high level of uncertainty. Figure 8 illustrates the actual and predicted customer demand, with 10th and 90th percentiles. In addition, there are other forms of uncertainty: the process of loaded cars becoming empty, the transit times, and the acceptability of a car to a shipper (is it clean enough?). There are five types of information processes that railroads have to deal with:

(1) Orders – Customers call in orders for cars, typically the week before when they are needed. The order does not include the destination of the order.
(2) Order destination – The destination of the order is not revealed until after the car is loaded.
(3) Empty cars – Empty cars become available from four potential sources: cars being emptied by shippers, empty cars coming on-line from other railroads, cars that have just been cleaned or repaired, and new cars that have just been purchased or leased.
(4) Transit times – As a car progresses through the network, we learn the time required for specific steps (after they are completed).

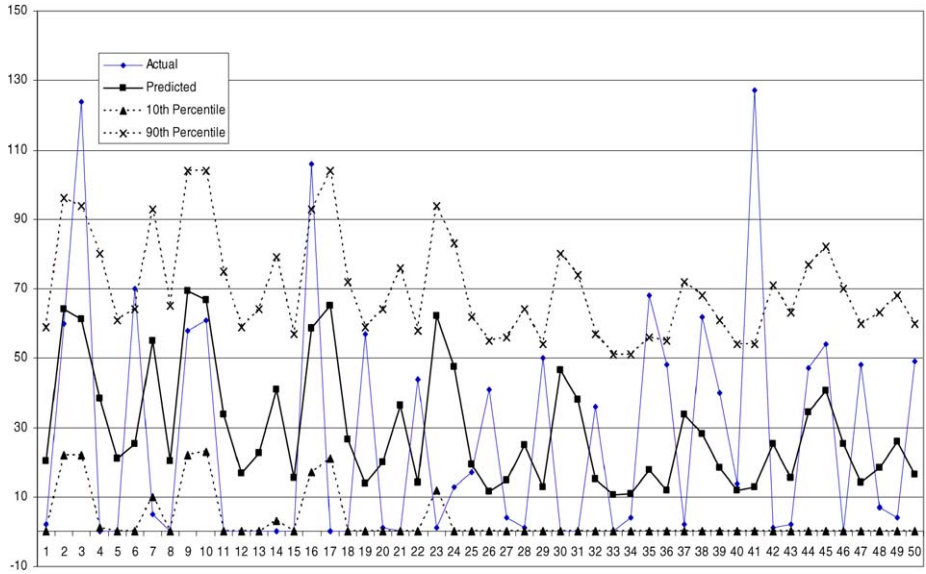


Fig. 8. Actual vs. predicted forecasts of future demands, showing the 10th and 90th percentiles.

(5) Updates to the status of a car – Cars break down ("bad order" in the language of railroads) or are judged (typically by the customer) to be not clean enough.

Railroads manage this uncertainty using several strategies. First, there is the fact that customers make their orders partially known in advance. For example, customers might book their orders the week before, although it is common for them to do so toward the end of the week. However, there can be discrepancies between what they book in the computer vs. what they really need (which might be communicated by phone). For example, if a railroad is running short on cars, customers have a tendency to overbook. Second, railroads depend on various forms of substitution when they cannot fill an order. Three forms of substitution come into play:

(1) Geographic substitution – The railroad may look at different sources of cars and choose a car that is farther away.
(2) Temporal substitution – The railroad may provide a car that arrives on a different day.
(3) Car type substitution – The railroad may try to satisfy the order using a slightly different car type.

It is common to model the decisions of moving cars from one location to another, but in practice car distribution is characterized by a number of decisions, including:

(1) Move car to a location – An empty car may be moved to a regional depot or an intermediate classification yard.
(2) Assign to a customer order – Here we make the explicit assignment of a car to a specific customer order.
(3) Clean or repair a car – This produces a change in the status of the car.
(4) Switch pools – Many cars belong to shipper pools which might be adjusted from time to time.
(5) Buy/sell/lease decisions – These decisions affect the overall size of the fleet.

Our presentation will consider only the first two classes, although classes 3 and 4 represent trivial extensions. To represent these decisions, we define

$\mathcal{D}^{\mathrm{c}}$ = the decision class to send cars to specific customers, where $\mathcal{D}^{\mathrm{c}}$ consists of the set of customers (each element of $\mathcal{D}^{\mathrm{c}}$ corresponds to a customer location),

$\mathcal{D}^{\mathrm{o}}$ = the decision to assign a car to a type of order. For $d \in \mathcal{D}^{\mathrm{o}}$, we let $b_d \in \mathcal{B}$ be the attributes of the order type associated with decision d,

$\mathcal{D}^{\mathrm{rd}}$ = the decision to send a car to a regional depot,

$\mathcal{D}^{\text{cl}}$ = the decision to send a car to a classification yard (each element of $\mathcal{D}^{\text{cl}}$ is a classification yard),

d^{ϕ} = the decision to hold the car ("do nothing").

The next step is to model the exogenous information processes. We start by modeling the arrivals of new resources, which includes both cars and orders. For example, a railroad might send a car "off line" to another railroad. When it returns to the railroad, it is as if a new car is arriving to the system. In some cases, a car sent to a shipper can be modeled as a car that is effectively disappearing from the network; when the car is released and returned from the shipper, it is a new arrival. New orders, of course, arrive in the form of phone calls or entries into a computer database (for example, through a web site).

We take advantage of our ability to separate the arrival process of information from the physical arrival of cars and orders. For example, we would represent new orders using:

$\widehat{R}^{\text{o}}_{tt'}$ = the vector of changes to the order vector (new customer orders, changes in orders) that arrive during time interval t that become actionable at time $t' \geqslant t$,

$\widehat{R}^{\text{o}}_{t} = \big(\widehat{R}^{\text{o}}_{tt'}\big)_{t' \geqslant t}$.

We would define similar vectors $\widehat{R}^{\text{c}}_{tt'}$ and $\widehat{R}^{\text{c}}_{t}$ for cars. Let W_t be our general variable representing the information arriving at time t, where

$$W_t = \big(\widehat{R}^{\text{o}}_{t}, \widehat{R}^{\text{c}}_{t}\big).$$

11.1.2 Optimization formulation

The most basic optimization model for car problems is the one that matches known cars to known orders over a predetermined horizon (even the most basic model captures the fact that some cars and orders are actionable in the future). This model can be stated as

$$\max_{x} \sum_{a \in \mathcal{A}} \sum_{d \in \mathcal{D}} c_{0ad} x_{0ad} \tag{81}$$

subject to

$$\sum_{d \in \mathcal{D}^{\text{c}}} x_{0ad} = R^{\text{c}}_{0a}, \quad a \in \mathcal{A}, \tag{82}$$

$$\sum_{a \in \mathcal{A}} x_{0ad} \leqslant R^{\text{o}}_{0b_d}, \quad d \in \mathcal{D}^{\text{o}}, \tag{83}$$

$$x_{0ad} \in Z_+, \quad a \in \mathcal{A}, d \in \mathcal{D}^{\text{c}}. \tag{84}$$

Equation (82) captures flow conservation constraints on cars only, while Equation (83) ensures that we do not assign more cars to orders than we have orders (effectively enforcing flow conservation on orders). This model follows

the normal convention of modeling cars as an *active* resource (resources that we actively modify) while orders are a passive resource (they only change their status if they are assigned to a car). Recall that in Equation (83), for every element of $d \in \mathcal{D}^{\mathrm{o}}$ there is an order *type* with attribute b_d.

The next step up in sophistication is a model that incorporates point forecasts of cars and orders that become known in the future (this model is in use by at least one railroad). Let $\overline{R}^{\mathrm{c}}_t$ and $\overline{R}^{\mathrm{o}}_t$ be point forecasts of cars and orders of information that will become available during time period t. This model can be written as

$$\max_x \sum_{a\in\mathcal{A}} \sum_{d\in\mathcal{D}} c_{0ad} x_{0ad} \tag{85}$$

subject to

$$\sum_{d\in\mathcal{D}} x_{0ad} = R^{\mathrm{c}}_{0a} + \sum_{t\in\mathcal{T}^{ph}\setminus 0} \overline{R}^{\mathrm{c}}_{ta}, \quad d \in \mathcal{D}^{\mathrm{c}}, a \in \mathcal{A}, \tag{86}$$

$$\sum_{a\in\mathcal{A}} x_{0ad} \leqslant R^{\mathrm{o}}_{ob_d} + \sum_{t\in\mathcal{T}^{ph}\setminus 0} \overline{R}^{\mathrm{o}}_{tb_d}, \quad d \in \mathcal{D}^{\mathrm{o}}, \tag{87}$$

$$x_{0ad} \in Z_+. \tag{88}$$

Note that this model aggregates forecasts of all future information about cars and orders. We retain the information about when the order will be actionable in the vector b_d, so that decisions to assign, say, a car available now to an order that is forecasted to arrive at time t' which will then be actionable at time t'', can properly account for the difference between the available time of the car and the available time of the order.

Another strategy is to use our dynamic programming approximations as we outlined above. If we use a (possibly nonlinear), separable value function, we would find ourselves solving

$$X^{\pi,n}_t(R_t) = \arg\max_{x_t}\left(c_t x_t + \sum_{t'\geqslant t} \sum_{a\in\mathcal{A}} \overline{V}^{n-1}_{t,t'a}\left(R^x_{t,t'a}(x_t)\right)\right) \tag{89}$$

subject to

$$\sum_{d\in\mathcal{D}} x_{tad} = R^{\mathrm{c}}_{t,ta}, \quad a \in \mathcal{A}, \tag{90}$$

$$\sum_{a\in\mathcal{A}} x_{tad} \leqslant R^{\mathrm{o}}_{tb_d}, \quad d \in \mathcal{D}^{\mathrm{o}}, \tag{91}$$

$$x_{tad} \in Z_+. \tag{92}$$

Equation (90) limits us to acting on cars that are actionable now, while Equation (91) says that we cannot move more cars loaded than the orders we know about now. This problem is generally quite easy to solve, but does require the iterative learning logic described in Section 9.

11.1.3 Some numerical experiments

The approximate dynamic programming algorithm was run on a dataset from a major railroad using nonlinear, separable value functions. The results are presented in Figure 9 which shows total profits, along with total revenues, empty movement costs, and service penalty costs. The figure shows a steady improvement in total profits as the objective function improves. The total revenue drops in the early iterations but rises back to approximately the same level as the first iteration, indicating that our improvement is through better repositioning rather than higher coverage. These experiments indicate that the adaptive learning logic works, but raises the question: exactly what is it doing that is smarter than a myopic model?

A common misconception is that we do not need to look into the future because customer orders are known in advance. For the rail dataset, customers orders are generally known the week before, with the bulk of customers calling in their orders on Wednesday and Thursday. Empty movement times are typically less than five days, which suggests that the prebook time allows the railroad to wait until orders are known before moving a car. Figure 10 shows the percent of total miles that are empty in history (approximately 54 percent for this dataset), from a myopic optimization model (which uses information about cars and orders that become available in the future), and from an approximate dynamic programming model. The myopic model reduces empty miles to approximately 48 percent, which is a significant reduction. Other railroads have used estimated savings such as these to justify projects in the millions of dol-

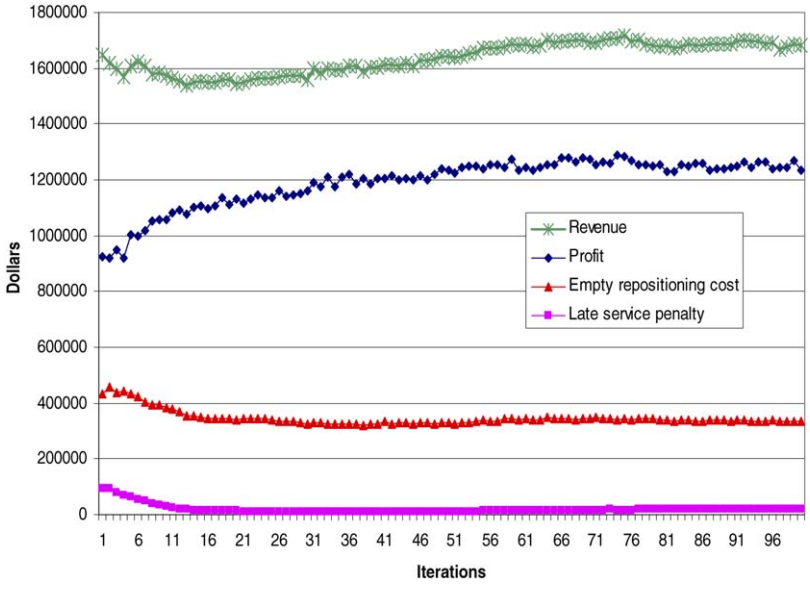


Fig. 9. The contribution of adaptive learning to revenue, costs and overall profits.

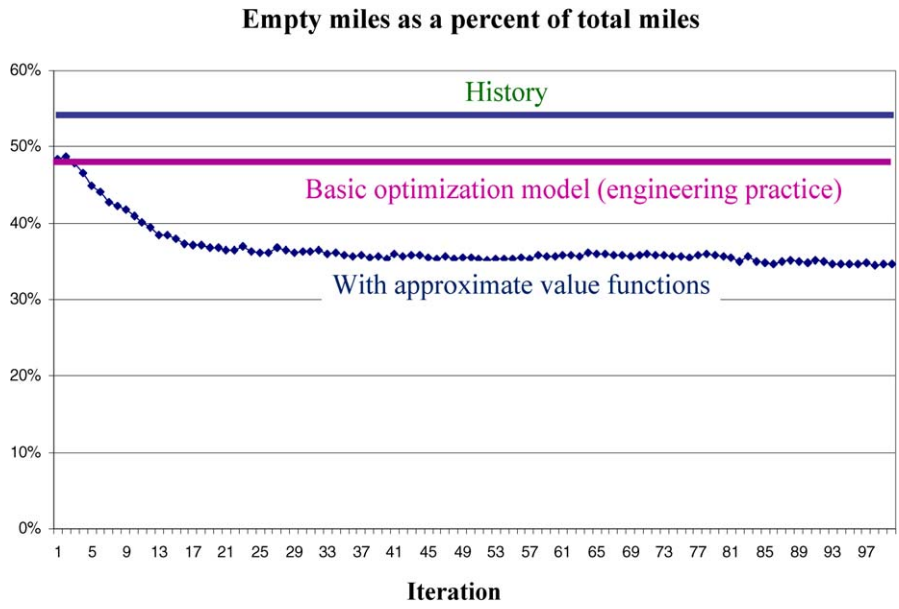


Fig. 10. Empty miles as a percent of total (a) in history, (b) optimized using a myopic model, and (c) optimized using an approximate dynamic programming procedure.

lars. The adaptive learning model reduces empties to about 35 percent after 25 iterations, with most reductions after this.

Where are these benefits from approximate dynamic programming coming from? There appear to be two sources. First, when a car becomes empty and available at a shipper, it *must* be moved; it cannot be held at the shipper. Cars that become empty early in the week often have to be moved before customer orders have been booked. A common strategy used by railroads is to move cars to classification yards and regional depots where they can be reassigned to customers after they arrive (by this time, customers orders will often have arrived). A myopic model will never send a car to these locations because there are no customer demands at these locations. Using approximate value functions, the model will send cars to locations where they have the highest value, which could easily be a regional depot or classification yard.

The second, and perhaps more significant reason that approximate dynamic programming works is that it captures the uncertain presence of cars in the future. In a snapshot of cars and orders, there is often far more information about orders in the future than cars. A customer might book his order a week or more into the future, but we are generally unable to track cars into the future. We cannot forecast the availability of cars after they fulfill orders that have not even been called in yet. Furthermore, even orders that have been called in are incomplete; shippers do not identify the destination for a car until after the car has been loaded.

The behavior that we have observed is that a myopic model has a tendency to assign a car that is available now to an order that is known now, but does not have to be served until the future. The myopic model is not able to assign a car after it has been used to fulfill an order that is either unknown or has incomplete information (e.g., the destination). The approximate dynamic programming model takes this information implicitly into account when it solves subproblems in the future. Thus, the adaptive learning technology is not only learning about future orders that are currently unknown, but also future cars.

11.1.4 Notes

The car distribution problem captures some of the richness of dynamic information processes that arise in real problems. At the same time, it offers some convenient simplifications. First, we implicitly assume that the attribute spaces $\mathcal{A}$ and $\mathcal{B}$ are fairly small and easy to enumerate. This is not too important in the myopic and rolling horizon models, but is very important if we use the dynamic programming approximation. The reason is that if we act on a car to produce a car with attribute a', we need to have a value function $\overline{V}^{n-1}_{t,t'a'}$ to measure the value of the resource. Furthermore, we have to update these value functions even if R_{ta} is zero, since an early estimate of the value of a car in the future might have been much lower than what it should be. If $\overline{V}^{n-1}_{t,t'a'}$ underestimates the value of a car of type a', then we might not implement a decision that produces this car type in the future again. As a result, we have to keep updating the value function to get a good estimate. This has proven to be very important in practice.

In addition, although this is an instance of a two-class resource allocation problem, it is modeled as a one-layer problem. We only model decisions acting on cars, and we only estimate the value of cars in the future (since an order, once it has been moved, is assumed to disappear).

11.2 The dynamic assignment problem for truckload trucking

We next consider the problem of assigning drivers to loads that arises in truckload trucking. We first present the model of the physical process, and then discuss methods for making decisions.

11.2.1 A model of the dynamic load assignment problem

Unlike our car distribution problem, we are going to model this as a two layer problem from the perspective of approximating the value of resources in the future. We also use this problem to illustrate some other modeling devices. For example, in the car distribution problem we represented car types and demand types. This is more natural since there are often a large number of cars (and even orders) of exactly the same type. In the dynamic assignment problem, we are modeling the assignment of "resources" to "tasks" which

are discrete in nature (and generally highly individualized). It is more natural, then, to model each resource and task individually. We note that we refer to these as two "resource layers" but use the more common term "resource" to refer to the layer that represents the objects (drivers, trucks, locomotives) and "task" to refer to the customer or requirement to be satisfied.

We begin by defining

$$\begin{aligned}
\mathcal{C}^{\mathrm{R}} &= \text{set of resource classes} \\
&= \big(\text{Drivers(D)}, \text{Loads(L)}\big), \\
\mathcal{R}^{c} &= \text{set of all resource indices for class } c \in \mathcal{C}^{\mathrm{R}} \text{ that might possibly} \\
&\qquad \text{enter the system,} \\
\mathcal{R} &= \mathcal{R}^{\mathrm{D}} \cup \mathcal{R}^{\mathrm{L}}.
\end{aligned}$$

We use d to index elements of the set $\mathcal{R}^{\mathrm{D}}$ and l to index elements of the set $\mathcal{R}^{\mathrm{L}}$, and we use $r \in \mathcal{R}$ as a generic resource which may be a driver or a load. This convention allows us to avoid defining everything twice.

To describe the state of our system, we define

$$\begin{aligned}
R_{t,t'r} &= \begin{cases} 1, & \text{if resource } r \in \mathcal{R} \text{ is known at time } t \text{ and available} \\ & \quad \text{to be assigned in period } t', \\ 0, & \text{otherwise,} \end{cases} \\
R_t &= (R_{t,t'r})_{r\in\mathcal{R},t'\geqslant t}.
\end{aligned}$$

We use R_t^{D} and R_t^{L} to refer specifically to the resource state vector for drivers and loads.

Over time, new drivers and loads will arrive to the system. We assume that there is a master list of identifiers for all drivers and loads (in the set $\mathcal{R}$), but that there is a specific time at which a driver or load becomes known. We model this process using

$$\begin{aligned}
\widehat{R}_{t,t'r} &= \begin{cases} 1, & \text{if resource } r \in \mathcal{R}, \text{ which is available to be acted on} \\ & \quad \text{at time } t', \text{ first becomes known in period } t, \\ 0, & \text{otherwise,} \end{cases} \\
\widehat{R}_t &= (\widehat{R}_{t,t'r})_{r\in\mathcal{R},t'\geqslant t}.
\end{aligned}$$

In this simple model, the exogenous information arriving in time period t is given by

$$W_t = \big(\widehat{R}_t^{\mathrm{D}}, \widehat{R}_t^{\mathrm{L}}\big). \tag{93}$$

Our decision problem involves the assignment of drivers to loads, which are represented using

$$\begin{aligned}
x_{tdl} &= \begin{cases} 1, & \text{if driver } d \text{ is assigned to load } l \text{ at time } t, \\ 0, & \text{otherwise,} \end{cases} \\
x_t &= (x_{tdl})_{d\in\mathcal{R}^{\mathrm{D}},l\in\mathcal{R}^{\mathrm{L}}},
\end{aligned}$$

$$x^{\mathrm{L}}_{tl} = \sum_{d \in \mathcal{R}^{\mathrm{D}}} x_{tdl}$$
$$= \begin{cases} 1, & \text{if any driver is assigned to load } l \text{ at time } t, \\ 0, & \text{otherwise,} \end{cases}$$
$$x^{\mathrm{D}}_{td} = \sum_{l \in \mathcal{R}^{\mathrm{L}}} x_{tdl}$$
$$= \begin{cases} 1, & \text{if driver } d \text{ is assigned to any load at time } t, \\ 0, & \text{otherwise.} \end{cases}$$

We use these variables to write our resource dynamics as

$$R^{\mathrm{D}}_{t+1} = R^{\mathrm{D}}_t - x^{\mathrm{D}}_t + \widehat{R}^{\mathrm{D}}_{t+1}, \tag{94}$$
$$R^{\mathrm{L}}_{t+1} = R^{\mathrm{L}}_t - x^{\mathrm{L}}_t + \widehat{R}^{\mathrm{L}}_{t+1}. \tag{95}$$

11.2.2 Optimization formulation

Our challenge, of course, is one of designing a decision function $X^{\pi}_t(R_t)$. Our goal is to maximize the total contribution from assigning drivers to loads over a planning horizon T^{ph}. Let

$$c_{tdl} = \text{the contribution from assigning driver } d \text{ to load } l \text{ at time } t.$$

Our one period contribution function is given by

$$C_t(x_t) = \sum_{r \in \mathcal{R}^{\mathrm{D}}_t} \sum_{l \in \mathcal{R}^{\mathrm{L}}_t} c_{tdl} x_{tdl}. \tag{96}$$

We assume that we only receive a contribution when a decision is implemented, which can only occur when both the resource and task are actionable. However, we can *plan* decisions in the future, which is what we assume we are doing when we assign a driver to a load when one or both is actionable in the future. As discussed in Section 7 we make decisions by solving a problem of the form

$$X^{\pi}_t(R_t) = \arg\max_x C^{\pi}_t(x). \tag{97}$$

Keep in mind that the function X^{π}_t returns a vector $x_t = (x_{tt'})_{t' \geqslant t}$, which means it might include actions to be implemented now, x_{tt}, and plans of what to do in the future, $(x_{tt'})_{t' \in \mathcal{T}^{ph} \setminus 0}$. We may limit our attention to resources that are knowable now and actionable now, R_{tt}, or we may include resources that are actionable in the future, $(R_{tt'})_{t' \in \mathcal{T}^{ph}}$. Both represent forms of deterministic models.

An interesting (and perhaps unexpected) behavior arises when we consider resources that are known now but actionable in the future. It would seem natural that such a model would outperform a model that focuses only on resources that are actionable now, but looking further into the future can produce poor results if it is not done well. In most applications, the extent to which

we know about internal resources (drivers, trucks, locomotives, freight cars) in the future can be quite different from how far we know about customer orders. In rail and trucking, it is typical for the customers to call in orders farther into the future than we would know about drivers and equipment. For example, it can be hard to track rail equipment into the future because even when a customer order is known, we might not know everything about the order (most obviously, the destination of the order, but also the time required to complete the order). This can make it difficult to track a car into the future for more than a few days. Yet customers will call in orders a week or even two weeks into the future.

Incorporating point forecasts of the future can be difficult for this problem class, since all the activities are (0/1). First, while we have made the explicit choice to model specific drivers and loads, as opposed to driver and load types (as we did in our car distribution model), when we perform forecasting we have to forecast by type (but we can treat a forecasted driver or load as if it were real). Generally, we are not able to forecast as many attributes of a real resource, which means that forecasted resources have to be somewhat simplified. Second, we encounter problems if we have to forecast a type of load that occurs with probability 0.20.

An alternative strategy for handling uncertain, forecasted activities is to resort to value function approximations. Here again, it is more useful to think of estimating the value of a type of driver or load, rather than a specific driver or load. To handle this transition, we let

$$a_{tr} = \text{the attributes of resource (driver or load) } r \text{ at time } t.$$

If we decide to hold a driver (not assign the driver to any load) then this means we have this same driver in the future. This is a "do nothing" decision (which we have represented using d^{ϕ}) which might produce a modified set of attributes given by $a^{\mathrm{M}}(t, a_r, d^{\phi})$. We note that

$$\begin{aligned} R_{td}^{x,\mathrm{D}} &= \bigl(1 - x_{td}^{\mathrm{D}}\bigr) \\ &= \begin{cases} 1, & \text{if driver } d \text{ is held at time } t, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Similarly, let $R_{tl}^{x,\mathrm{L}}$ indicate whether load l is held until the next time period. $R_t^x = (R_t^{x,\mathrm{D}}, R_t^{x,\mathrm{L}})$ is our post-decision resource state variable. The decision function is given by

$$X_t^{\pi}(R_t) = \arg\max_{x \in \mathcal{X}_t} \sum_{d \in \mathcal{R}_t^{\mathrm{D}}} \sum_{l \in \mathcal{R}_t^{\mathrm{L}}} \bigl(c_{tdl} x_{tdl} + \overline{V}_t\bigl(R_t^x\bigr)\bigr) \tag{98}$$

subject to appropriate flow conservation constraints on drivers and loads. We then have to determine an appropriate functional form for $\overline{V}_t(R_t)$. The most

obvious to choose is a linear approximation

$$\overline{V}_t(R_t) = \sum_{d \in \mathcal{R}_t^{\mathrm{D}}} \bar{v}_{td}^{\mathrm{D}} R_{td}^{\mathrm{D}} + \sum_{l \in \mathcal{R}_t^{\mathrm{L}}} \bar{v}_{tl}^{\mathrm{L}} R_{tl}^{\mathrm{L}}. \tag{99}$$

Here $\bar{v}_{td}^{\mathrm{D}}$ is the value of driver d if it is held at time t, and $\bar{v}_{tl}^{\mathrm{L}}$ is the value of load l. We note that this approximation is, for the first time, capturing the value of holding both resource layers in the future. We can easily estimate these values directly from the dual variables for the driver and load resource constraints (94) and (95). One complication is that the dual variables, since they are simply subgradients, are typically not accurate enough. If $R_{td}^{x,\mathrm{D}} = 1$, then we want the impact of eliminating driver d from the future; if $R_{td}^{x,\mathrm{D}} = 0$, then we want the value of adding driver d to the future. It is possible to find these left and right gradients (respectively) by solving flow augmenting path problems into and out of the supersink to all the nodes in the network.

If we introduce our linear approximation into (98), we obtain, after some manipulation

$$X_t^{\pi}(R_t) = \arg\max_{x \in \mathcal{X}_t} \sum_{d \in \mathcal{R}_t^{\mathrm{D}}} \sum_{l \in \mathcal{R}_t^{\mathrm{L}}} \left(c_{tdl} - \bar{v}_{t,a^{\mathrm{M}}(t,a_d,d^{\phi})}^{\mathrm{D}} - \bar{v}_{t,a^{\mathrm{M}}(t,a_l,d^{\phi})}^{\mathrm{L}}\right) x_{tdl}, \tag{100}$$

where $a^{\mathrm{M}}(t, a_d, d^{\phi})$ is the attribute vector of a driver that has been held, and $a^{\mathrm{M}}(t, a_l, d^{\phi})$ is the attributes of a load. We note that (100) is still an assignment problem which is an important practical result.

There is still a problem. If $x_{tdl} = 1$, then we are assigning driver d to load l, and *both* are removed from the problem in the future. So, each decision at time t impacts a driver *and* a load. If x_{tdl} is increased from 0 to 1, then we have the effect of decreasing R_{td}^{D} and R_{tl}^{L} each by one. This is approximated by $-(\bar{v}_{td}^{\mathrm{D}} + \bar{v}_{tl}^{\mathrm{L}})$. Needless to say, this will introduce errors. A better approximation would be to capture the nonseparability of the value function. We could, instead, let

$$\bar{v}_{tdl} = \text{the marginal contribution of holding driver } d \text{ and load } l \text{ at the same time.}$$

Of course, we use $-\bar{v}_{tdl}$ to approximate the impact of assigning driver d to load l and therefore eliminating both of them from the pool of resources in the future. We now find ourselves solving subproblems of the form

$$X_t^{\pi}(R_t) = \arg\max_{x \in \mathcal{X}_t} \sum_{d \in \mathcal{R}_t^{\mathrm{D}}} \sum_{l \in \mathcal{R}_t^{\mathrm{L}}} (c_{tdl} - \bar{v}_{tdl}) x_{tdl}. \tag{101}$$

As before, (101) is still also an assignment problem, which means that if we can estimate $\bar{v}_{tdl}$, then our single-period problem in Equation (98) is no more difficult than the original myopic problem.

Not surprisingly, $\bar{v}_{tdl}$ (see Spivey and Powell, 2004, for details) is considerably harder to compute than $\bar{v}^{\mathrm{D}}_{td}$ and $\bar{v}^{\mathrm{L}}_{tl}$, which require only two flow-augmenting path calculations. $\bar{v}_{tdl}$, by contrast, requires a flow augmenting from each load node l back to each driver node d. This can be accomplished with a flow-augmenting path for each driver, which is substantially more difficult than the single flow augmenting path into and out of the supersink we required for the separable approximation.

11.2.3 Some numerical experiments

The optimization model allows us to provide answers (experimentally) to two types of questions. First, does an ADP policy (that is, a policy that uses approximate dynamic programming) provide solutions that are better than myopic policies, and how does this compare with the availability of advance information? Second, what is the value of advance information, and how does the value of advance information change with the sophistication of the policy used? These two questions illustrate the two perspectives of dynamic models that we addressed in the Introduction.

These questions were addressed in Spivey and Powell (2004). The issue of solution quality was addressed by running a series of simulations. For each simulation, we can compute the optimal solution after all the information is known (this is simply an assignment problem using all the resources and tasks), giving us a posterior bound. The first set of simulations used deterministic data, which is to say that $|\Omega| = 1$. For this case, we can hope to achieve a solution that is close to the posterior bound. The second set of simulations used a stochastic data set, with $|\Omega| \gg 1$, where there are a number of different outcomes of new drivers and loads.

The results of the simulations are summarized in Figure 11, which represent an average over 10 different datasets. For the deterministic datasets, we see a steady improvement as we move from a myopic policy ($\bar{v}^{\mathrm{R}} = \bar{v}^{\mathrm{L}} = 0$), to resource gradients (where we use only $\bar{v}^{\mathrm{R}}$), to resource and task gradients (where we use $\bar{v}^{\mathrm{R}}$ and $\bar{v}^{\mathrm{L}}$), to the arc gradients (where we use $\bar{v}_{tdl}$). It is very encouraging that the arc gradients give near optimal solutions. For stochastic datasets, the best results are obtained when we use resource and task gradients; interestingly, the far more computationally intensive arc gradients do not seem to add any value. It appears that the added precision that these gradients bring are lost in the presence of uncertainty.

It is usually the case in transportation that we will know about resources and tasks in the future, represented by the resource vector $R_{tt'}$. Simulations were run which allowed the decision function to consider resources and tasks that are known now (at time t) but actionable in the future (at time t') as long as the actionable time is within a planning horizon T^{ph}. We would expect to get better results as the planning horizon is increased (which is an indication of the value of advance information) but it is not clear how much this answer is influenced by the use of value function approximations.

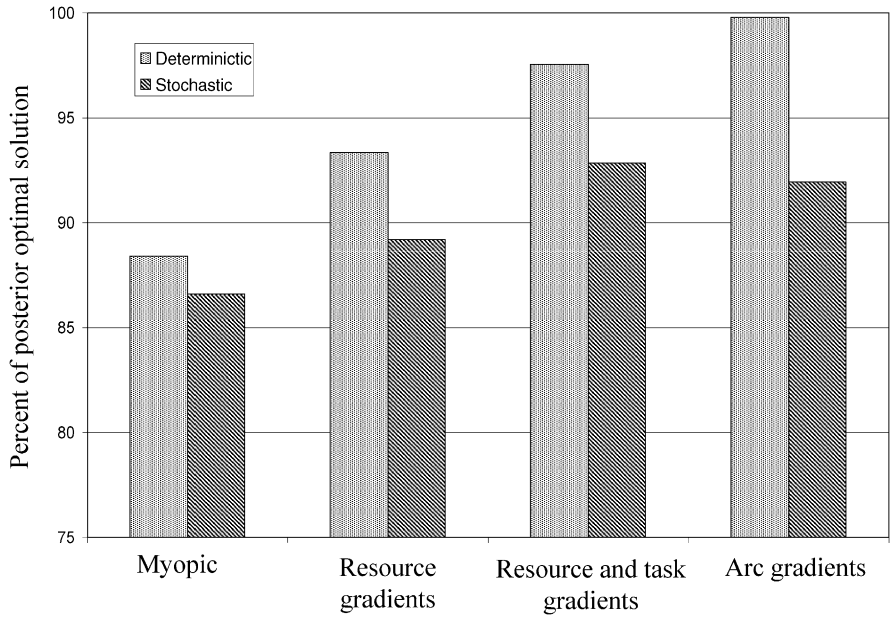


Fig. 11. The effect of resource gradients alone, resource and task gradients, and arc gradients, on solution quality for deterministic and stochastic problems.

Figure 12 shows the improvement in total profits (always measured against the posterior bound) for both myopic and ADP policies. In each case, the policies were tested not only for different values of the planning horizon T^{ph} (the time range over which information is used) but also for different decision horizons (the time period over which decisions in the future are locked into place). The results show that increasing the planning horizon is significant in the case of a myopic policy, but is surprisingly almost nonexistent when we use an ADP policy. At a minimum, it appears that ADP policies benefit less from advance information than a myopic policy. The implication is that measuring the value of advance information using myopic policies might produce inflated estimates.

11.3 Batch processes for less-than-truckload trucking

Up to now, we have considered problems that can be viewed as "flow problems". Although both the car distribution problem and load matching problem involve discrete resources, these are naturally formed as linear programs with integrality requirements, and they can either be solved as pure networks or linear programs which provide near-integer solutions.

A problem class where this is not the case arises in batch processes such as less-than-truckload trucking, where we have to consolidate smaller packages onto larger vehicles. The decision is how to route the shipments, and when and where to send the trucks. The decision to send a truck has to balance the value

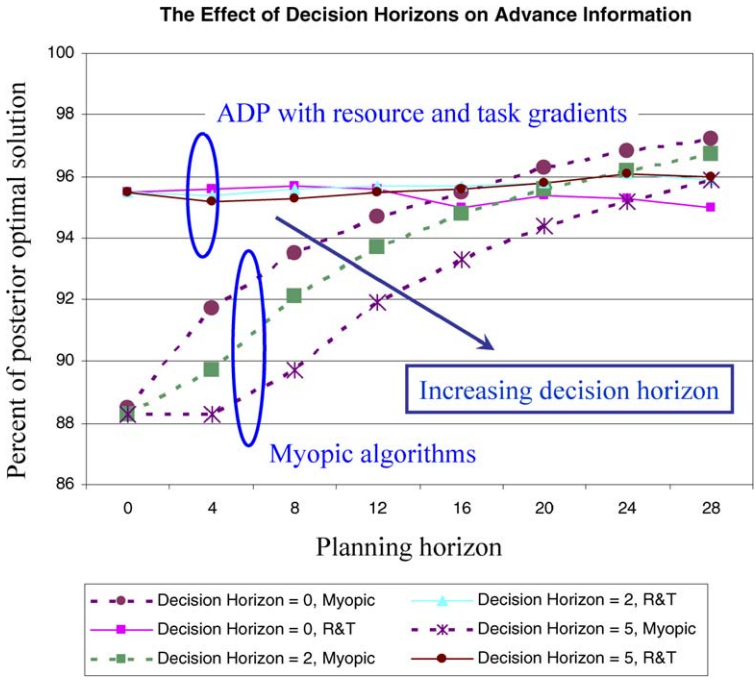


Fig. 12. The value of advance information as estimated by a myopic policy, and using approximate dynamic programming (from Spivey and Powell (2004)).

of moving the shipments now against various costs of holding the shipments. Since sending a truck has the same mathematical structure as building a link in the network, this has been referred to as the dynamic network design problem (Crainic, 2000).

As of this writing, we are not aware of any computationally tractable algorithms that have proven successful for realistic instances of the dynamic network design problem, at least as it arises in less-than-truckload trucking. Heuristics have been applied with varying degrees of success for static instances (Powell, 1986; Crainic and Roy, 1988; Crainic and Rousseau, 1988), but the dynamic versions are much harder. Most of the research has focused on solving the problem with a single link. If we face the problem of dispatching a truck over a single link with homogeneous customers, then we have a textbook dynamic program that is easy to solve.

The challenge arises when there is more than one product type. Speranza and Ukovich (1994, 1996) develop optimal strategies for the deterministic multiproduct problem. Bertazzi and Speranza (1999a, 1999b) consider the deterministic problem of shipping products from an origin to a destination through one or several intermediate nodes and compare several classes of heuristics including decomposition of the sequence of links, an EOQ-type solution and a dynamic programming-based heuristic. The results assume that demands are

both deterministic and constant. Bertazzi et al. (2000) consider a stochastic, multiproduct problem which is solved using approximate dynamic programming techniques, but the method is only tested with a single product type.

In this section, we briefly review the single-link model described in Papadaki and Powell (2003), which considers multiple product types (the method is tested on a problem with 100 product types), uncertainty and significant nonstationarities. For transportation, a "product type" might refer to the destination of the shipment, the time at which the shipment has to arrive at the destination, and the shipment priority (other attributes include size and weight).

11.3.1 A single link model of the dynamic dispatch problem

We begin by defining the following:

Problem parameters:

$\mathcal{K}$ = set of customer classes,

c^{d} = cost to dispatch a vehicle,

c_k^{h} = holding cost of class $k \in \mathcal{K}$ per time period per unit product,

$c^{\mathrm{h}} = \left(c_1^{\mathrm{h}}, c_2^{\mathrm{h}}, \ldots, c_{|\mathcal{K}|}^{\mathrm{h}}\right)$,

K = service capacity of the vehicle, giving the total number of customers who can be served in a single dispatch.

Activity variables:

R_{tk}^x = number of customers in class k waiting at time t before new arrivals have been added (the post-decision state variable),

$R_t^x = \left(R_{tk}^x\right)_{k\in\mathcal{K}}$,

$\widehat{R}_t$ = vector random variable giving the number of arrivals in time t of each type of customer,

$R_t = R_{t-1}^x + \widehat{R}_t$

= number of customers waiting just before we make a decision at time t.

Decision variables:

x_{tk} = the number of customers in class k who are served at time t,

$X_t^\pi(R_t)$ = decision function returning x_t.

Our feasible region $\mathcal{X}_t$ is defined by:

$$x_{tk} \leqslant R_{tk},$$

$$\sum_{k\in\mathcal{K}} \leqslant K.$$

It is useful to define a dispatch function:

$$Z_t(x_t) = \begin{cases} 1, & \text{if } \sum_{k \in \mathcal{K}} x_{tk} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Written this way, $x_t = X_t^\pi$ determines Z_t. Later we show that we can actually determine Z_t first, and then compute X_t^π.

Objective function:
The one-period cost function is given by

$$C_t(R_t, \widehat{R}_t, x_t) = c^{\mathrm{d}} Z_t(x_t) + c^{\mathrm{h}}(R_t - x_t).$$

The objective function over the planning horizon can now be written

$$F(R_0) = \min_{\pi \in \Pi} E\left\{ \sum_{t=0}^{T} C_t\big(R_t, X_t^\pi(R_t)\big) \right\}.$$

11.3.2 A solution algorithm for the single-link problem

The most common practical approach for this problem is to use a simple dispatch rule that might look like "if we have at least a certain amount of freight, send the vehicle, otherwise hold it". Such rules generally have to be time dependent ("if it is Friday night…") or state dependent ("if the truck has been held more than 12 hours and it is at least 80 percent full…") since the arrival process of freight can be highly nonstationary. Any effort to use optimization has to recognize that such rules can be fairly effective, and at a minimum are easy to understand.

The value of solving this problem with optimization is to impose the discipline of formulating and minimizing a cost function. The problem is that algorithms for network design have generally not been very effective. We propose to solve this problem (which is high-dimensional because of the multiple product types) using approximate dynamic programming.

Using our standard methodology, we can formulate a decision function using

$$X_t^\pi(R_t) = \arg\min_{x_t} \big(C_t(R_t, x_t) + \overline{V}_{t+1}^n\big(R_t(\omega_t), x_t\big) \big). \tag{102}$$

It is possible to show that the value function increases monotonically in the resources; in effect, the more customers that are waiting, the higher the cost. The function itself is neither concave or convex, but especially when we introduce uncertainty, it is approximately linear, suggesting the approximation

$$\overline{V}_t(R_t) = \bar{v}_t R_t.$$

Before, we estimated our slopes using dual variables from solving the subproblems. In the case of batch processes, we do not have access to dual variables, but we can use finite differences

$$\tilde{v}_{kt} = \widetilde{V}_t(R_t + e_k, \omega) - \widetilde{V}_t(R_t, \omega),$$

where e_k is a $|\mathcal{K}|$-dimensional vector with a single 1 in the kth element. Performing these for each product class can be extremely expensive. A short-cut, suggested in Graves (1981) in the context of multistage lot-sizing, is to assume that increasing the number of customers (of any type) does not change the dispatch decision at time t. This makes it trivial to determine the marginal impact of extra customers.

Rather than solve (102) to find x_t, it actually makes more sense to find z_t first (that is, determine whether the batch will be sent or held), and then compute x_t using a simple rule. If $z_t = 0$, then clearly $x_t = 0$. If $z_t = 1$ and $\sum_{k\in\mathcal{K}} R_{tk} \leqslant K$, then $x_{tk} = R_{tk}$. The only interesting case arises when $z_t = 1$ and $\sum_{k\in\mathcal{K}} R_{tk} > K$. Assume that the customer classes are ordered so that $c_1^{\mathrm{h}} \leqslant c_2^{\mathrm{h}} \leqslant \cdots \leqslant c_k^{\mathrm{h}}$. Then, we want to make sure customers in class 1 are all added to the batch before trying to add customers of class 2, and so on, until we reach the capacity of the batch. Assuming we use this rule, the steps of an approximate dynamic programming algorithm are summarized in Figure 13.

The algorithm was tested on a single-product problem in Papadaki and Powell (2003) so that comparisons against an optimal solution could be made (using classical dynamic programming). The approximate dynamic programming approach was compared to a rule which specified that the vehicle should be dispatched when full, or if it had been held τ units of time. For each dataset, τ was chosen so that it produced the best results (the assumption

Step 1. Given R_0: Set $\bar{v}_t^0 = 0$ for all t. Set $n = 1$, $t = 0$.
Step 2. Set $R_0^n = R_0$ and choose random sample ω^n.
Step 3. Calculate

$$z_t^n = \arg\min_{z_t\in\{0,1\}} \bigl\{c^{\mathrm{d}} z_t + c^{\mathrm{h}} \cdot \bigl(R_t^n - z_t X\bigl(R_t^n\bigr)\bigr) + \bar{v}_t^n\bigl(R_t^n - z_t X\bigl(R_t^n\bigr)\bigr)\bigr\}$$

and

$$R_{t+1}^n = R_t^n - z_t X\bigl(R_t^n\bigr) + \widehat{R}_{t+1}\bigl(\omega^n\bigr).$$

Then define:

$$\widetilde{V}_t^n\bigl(R_t^n\bigr) = \min_{z_t\in\{0,1\}} \bigl\{c^{\mathrm{d}} z_t + c^{\mathrm{h}} \cdot \bigl(R_t^n - z_t X\bigl(R_t^n\bigr)\bigr) + \bar{v}_t^n\bigl(R_t^n - z_t X\bigl(R_t^n\bigr)\bigr)\bigr\}.$$

Step 4. Update the approximation as follows. For each $k = 1, \ldots, |\mathcal{K}|$, let:

$$\hat{v}_{tk}^n = \widetilde{V}_t^n\bigl(R_t^n + e_k\bigr) - \widetilde{V}_t^n\bigl(R_t^n\bigr),$$

where e_k is a $|\mathcal{K}|$-dimensional vector with 1 in the kth entry and the rest zero. Update the approximation by smoothing:

$$\bar{v}_t^n = \bigl(1 - \alpha^n\bigr)\bar{v}_t^{n-1} + \alpha^n \hat{v}_t^n$$

Fig. 13. Approximate dynamic programming algorithm for the batch dispatch problem.

being that for any specific problem class, we could choose the best holding time).

The method was tested on a variety of datasets which included both stationary and nonstationary arrival processes. An important determinant of performance is the relative size of the holding cost per unit c^h, and the dispatch cost per unit of capacity c^d/K. For this reason, the datasets were divided into three groups: $c^h > c^d/K$, $c^h \simeq c^d/K$, and $c^h < c^d/K$. In the first group, the holding cost is high enough that we will tend to dispatch vehicles very quickly. In the last group, the holding cost is low enough that we expect a dispatch-when-full policy to be best. The middle group is the most interesting.

For each group, we further divided the runs between datasets where the arrival process followed a periodic (e.g., daily) pattern (which was highly nonstationary) from datasets where the arrival process was stationary. The results are reported in Table 3, which shows that the approximate dynamic programming approach works better than the optimized myopic heuristic, even for the relatively easier case $c^h < c^d/K$ where a dispatch-when-full strategy tends to work well (this strategy still struggles when the arrival process of customers is highly nonstationary).

Table 4 summarizes the results of runs done with 100 product types. In this case, we cannot solve the problem optimally, but we are able to compare against our myopic policy, and we have an idea from the single product case of how well the myopic policy works. In this table, instead of reporting the percent over optimal, we report the total costs of the approximate dynamic programming policy divided by the total costs of the myopic policy. If we again focus on the class of problems, where $h < c/K$ (where the myopic policy will work the best) we find that approximate dynamic programming strategies are approximately three percent better than the myopic policy, which is very consistent with our results from the single product case.

Table 3.
Fraction of total costs produced by each algorithm over the optimal cost: averages and standard deviations within each group (from Papadaki and Powell, 2003)

	Method: Iterations:	Linear (25)	Linear (50)	Linear (100)	Linear (200)	DWF-TC
$h > c/K$	Periodic	0.082	0.070	0.058	0.062	0.856
	Stationary	0.071	0.050	0.045	0.038	0.691
$h \simeq c/K$	Periodic	0.040	0.031	0.024	0.024	0.270
	Stationary	0.057	0.035	0.023	0.024	0.195
$h < c/K$	Periodic	0.029	0.025	0.019	0.019	0.067
	Stationary	0.031	0.019	0.015	0.013	0.059
Average		0.052	0.038	0.031	0.030	0.356

Table 4.
The (expected) cost of the approximate dynamic programming algorithm as a fraction of the cost of the DWF-TC myopic heuristic for both scalar (single product) and multiple product problems (from Papadaki and Powell, 2003)

	Method:	adp scalar	adp scalar	adp scalar	adp scalar	adp mult.	adp mult.	adp mult.	adp mult.
	Iterations:	(25)	(50)	(100)	(200)	(25)	(50)	(100)	(200)
$h > c/K$	Periodic	0.602	0.597	0.591	0.592	0.633	0.626	0.619	0.619
	Stationary	0.655	0.642	0.639	0.635	0.668	0.660	0.654	0.650
$h \simeq c/K$	Periodic	0.822	0.815	0.809	0.809	0.850	0.839	0.835	0.835
	Stationary	0.891	0.873	0.863	0.863	0.909	0.893	0.883	0.881
$h < c/K$	Periodic	0.966	0.962	0.957	0.956	0.977	0.968	0.965	0.964
	Stationary	0.976	0.964	0.960	0.959	0.985	0.976	0.971	0.969
Average		0.819	0.809	0.803	0.802	0.837	0.827	0.821	0.820

12 Perspectives on real-time problems

Dynamic models have a number of applications. We can simulate a dynamic process to better understand how to operate a system under more realistic settings. We might want to investigate the value of better information (through automated detection systems, better databases, processes that require customers to make requests known further into the future). A company might want to produce short term tactical forecasts (for example, to help identify bottlenecks) one or two days into the future given the state of the system right now.

One application that often arises at many companies is a desire to develop models to help run their operations in real time. While there are a number of ways to help a company make better decisions now (for example, with short term tactical forecasts), there is often an interest in having computers tell dispatchers and planners what to do (that is, to automate the decision itself).

There is an important, qualitative difference between a real-time, dynamic model that is used to look into the future to produce forecasts, and a real-time, dynamic model that is used to make specific, actionable recommendations. The first is trying to forecast activities in the future to help a human make better decisions now. The second, which is much harder, is actually trying to tell a human what to do right now. True real-time optimization can be viewed as the information-age version of factory robots. We anticipate that these will steadily make their way into operations, but adoption will be slow. Sometimes it is only when we try to replace a human that we fully appreciate the diversity of tasks that people perform.

References

Aronson, J., Chen, B. (1986). A forward network simplex algorithm for solving multiperiod network flow problems. *Naval Research Logistics Quarterly* 33 (3), 445–467.

Aronson, J., Thompson, G.L. (1984). A survey on forward methods in mathematical programming. *Large Scale Systems* 7, 1–16.

Assad, A. (1978). Multicommodity network flows: A survey. *Networks* 8 (1), 37–91.

Baker, S., Morton, D., Rosenthal, R., Williams, L. (2002). Optimizing military airlift. *Operations Research* 50 (4), 582–602.

Barnhart, C., Hane, C.A., Johnson, E.L., Sigismondi, G. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46 (3), 316–329.

Bellman, R. (1957). *Dynamic Programming*. Princeton Univ. Press, Princeton.

Bellman, R., Dreyfus, S. (1959). Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation* 13, 247–251.

Bertazzi, L., Speranza, M.G. (1999a). Inventory control on sequences of links with given transportation frequencies. *International Journal of Production Economics* 59, 261–270.

Bertazzi, L., Speranza, M.G. (1999b). Minimizing logistic costs in multistage supply chains. *Naval Research Logistics* 46, 399–417.

Bertazzi, L., Bertsekas, D., Speranza, M.G. (2000). Optimal and neuro-dynamic programming solutions for a stochastic inventory trasportation problem. Unpublished technical report, Universita Degli Studi Di Brescia.

Bertsekas, D., Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

Birge, J., Louveaux, F. (1997). *Introduction to Stochastic Programming*. Springer-Verlag, New York.

Blum, J. (1954). Multidimensional stochastic approximation methods. *Annals of Mathematical Statistics* 25, 737–744.

Cheung, R., Powell, W.B. (1996). An algorithm for multistage dynamic networks with random arc capacities, with an application to dynamic fleet management. *Operations Research* 44 (6), 951–963.

Cheung, R.K.-M., Powell, W.B. (2000). SHAPE: A stochastic hybrid approximation procedure for two-stage stochastic programs. *Operations Research* 48 (1), 73–79.

Cinlar, E. (2003). Private communication.

Crainic, T. (2000). Network design in freight transportation. *European Journal of Operational Research* 12 (2), 272–288.

Crainic, T., Rousseau, J.-M. (1988). Multicommodity, multimode freight transportation: A general modeling and algorithmic framework for the service network design problem. *Transportation Research B* 20, 290–297.

Crainic, T., Roy, J. (1988). OR tools for the tactical planning of freight transportation. *European Journal of Operations Research* 33, 290–297.

Desrochers, M., Soumis, F. (1989). A column generation approach to the urban transit crew scheduling problem. *Transportation Science* 23, 1–13.

Desrosiers, J., Soumis, F., Desrochers, M. (1984). Routing with time windows by column generation. *Networks* 14, 545–565.

Desrosiers, J., Solomon, M., Soumis, F. (1995). Time constrained routing and scheduling. In: Monma, C., Magnanti, T., Ball, M. (Eds.), *Network Routing*. *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, pp. 35–139.

Ford, L., Fulkerson, D. (1962). *Flows in Networks*. Princeton Univ. Press, Princeton.

Glockner, G.D., Nemhauser, G.L. (2000). A dynamic network flow problem with uncertain arc capacities: Formulation and problem structure. *Operations Research* 48, 233–242.

Glover, F., Karney, D., Klingman, D., Napier, A. (1974). A computation study on start procedures, basis change criteria, and solution algorithms for transportation problems. *Management Science* 20, 793–813.

Glover, F., Klingman, D., Phillips, N.V. (1992). *Network Models in Optimization and Their Application in Practice*. Wiley, New York.

Godfrey, G.A., Powell, W.B. (2001). An adaptive, distribution-free approximation for the newsvendor problem with censored demands, with applications to inventory and distribution problems. *Management Science* 47 (8), 1101–1112.

Godfrey, G.A., Powell, W.B. (2002). An adaptive, dynamic programming algorithm for stochastic resource allocation problems I: Single period travel times. *Transportation Science* 36 (1), 21–39.

Graves, S.C. (1981). Multi-stage lot sizing: An iterative procedure. In: Schwarz, L.B. (Ed.), *Multi-Level Production/Inventory Control Systems: Theory and Practice*. *TIMS Studies in the Management Sciences*, vol. 16. North-Holland, New York, pp. 95–110.

Infanger, G. (1994). *Planning under Uncertainty: Solving Large-Scale Stochastic Linear Programs*. *The Scientific Press Series*. Boyd & Fraser, New York.

Kall, P., Wallace, S. (1984). *Stochastic Programming*. Wiley, New York.

Kennington, J.L. (1978). A survey of linear cost multicommodity network flows. *Operations Research* 26, 209–236.

Kennington, J., Helgason, R. (1980). *Algorithms for Network Programming*. Wiley, New York.

Kushner, H.J., Yin, G.G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York.

Langley, R.W., Kennington, J.L., Shetty, C.M. (1974). Efficient computational devices for the capacitated transportation problem. *Naval Research Logistics Quarterly* 21, 637–647.

Lasdon, L. (1970). *Optimization Theory for Large Systems*. MacMillan Co, New York.

Lavoie, S., Minoux, M., Odier, E. (1988). A new approach of crew pairing problems by column generation and application to air transport. *European Journal of Operational Research* 35, 45–58.

Marar, A., Powell, W.B. (2004). Using static flow patterns in time-staged resource allocation problems. Technical report, Princeton University, Department of Operations Research and Financial Engineering.

Marar, A., Powell, W.B., Kulkarni, S. (2006). Combining cost-based and rule-based knowledge in complex resource allocation problems. *IIE Transactions* 38 (2), 159–172.

Morton, D.P., Rosenthal, R.E., Lim, T.W. (1996). Optimization modeling for airlift mobility. *Military Operations Research* 1, 49–67.

Morton, D.P., Salmeron, J., Wood, R.K. (2003). A stochastic program for optimizing military sealift subject to attack. *Stochastic Programming e-print Series*. Available at http://www.speps.info.

Papadaki, K., Powell, W.B. (2003). An adaptive dynamic programming algorithm for a stochastic multiproduct batch dispatch problem. *Naval Research Logistics* 50 (7), 742–769.

Powell, W.B. (1986). A local improvement heuristic for the design of less-than-truckload motor carrier networks. *Transportation Science* 20 (4), 246–257.

Powell, W.B. (1987). An operational planning model for the dynamic vehicle allocation problem with uncertain demands. *Transportation Research* 21, 217–232.

Powell, W.B., Shapiro, J.A., Simão, H.P. (2001). A representational paradigm for dynamic resource transformation problems. In: Coullard, R.F.C., Owens, J.H. (Eds.), *Annals of Operations Research*. J.C. Baltzer AG, Basel, pp. 231–279.

Powell, W.B., Shapiro, J.A., Simão, H.P. (2002). An adaptive dynamic programming algorithm for the heterogeneous resource allocation problem. *Transportation Science* 36 (2), 231–249.

Powell, W.B., Wu, T.T., Whisman, A. (2004). Using low dimensional patterns in optimizing simulators: An illustration for the airlift mobility problem. *Mathematical and Computer Modelling* 39, 657–675.

Puterman, M.L. (1994). *Markov Decision Processes*. Wiley, New York.

Robbins, H., Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics* 22, 400–407.

Sen, S., Higle, J. (1999). An introductory tutorial on stochastic linear programming models. *Interfaces* 29 (2), 33–61.

Shapiro, J., Powell, W.B. (2006). A metastrategy for dynamic resource management problems based on informational decomposition. *Informs Journal on Computing* 18 (1), 43–60.

Speranza, M., Ukovich, W. (1994). Minimizing trasportation and inventory costs for several products on a single link. *Operations Research* 42, 879–894.

Speranza, M., Ukovich, W. (1996). An algorithm for optimal shipments with given frequencies. *Naval Research Logistics* 43, 655–671.

Spivey, M., Powell, W.B. (2004). The dynamic assignment problem. *Transportation Science* 38 (4), 399–419.

Sutton, R. (1988). Learning to predict by the methods of temporal differences. *Machine Learning* 3, 9–44.

Sutton, R., Barto, A. (1998). *Reinforcement Learning*. MIT Press, Cambridge, Massachusetts.

Topaloglu, H., Powell, W.B. (2006). Dynamic programming approximations for stochastic, time-staged integer multicommodity flow problems. *Informs Journal on Computing* 18 (1), 31–42.

Tsitsiklis, J., Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control* 42, 674–690.

Van Roy, B. (2001). Neuro-dynamic programming: Overview and recent trends. In: Feinberg, E., Shwartz, A. (Eds.), *Handbook of Markov Decision Processes: Methods and Applications*. Kluwer Academic, Boston, pp. 431–460.

Vance, P.H., Barnhart, C., Johnson, E.L., Nemhauser, G.L. (1997). Airline crew scheduling: A new formulation and decomposition algorithm. *Operations Research* 45 (2), 188–200.

White, W. (1972). Dynamic transshipment networks: An algorithm and its application to the distribution of empty containers. *Networks* 2 (3), 211–236.

Wu, T.T., Powell, W.B., Whisman, A. (2003). The optimizing simulator: An intelligent analysis tool for the airlift mobility problem. Technical report, Princeton University, Department of Operations Research and Financial Engineering.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14006-2

Chapter 6

Vehicle Routing

Jean-François Cordeau

Canada Research Chair in Logistics and Transportation, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, H3T 2A7, Canada
E-mail: Jean-Francois.Cordeau@hec.ca

Gilbert Laporte

Canada Research Chair in Distribution Management, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, H3T 2A7, Canada
E-mail: gilbert@crt.umontreal.ca

Martin W.P. Savelsbergh

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA
E-mail: mwps@isye.gatech.edu

Daniele Vigo

Dipartimento di Elettronica, Informatica e Sistemistica, University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy
E-mail: dvigo@deis.umibo.it

1 Introduction

The *vehicle routing problem* lies at the heart of distribution management. It is faced each day by thousands of companies and organizations engaged in the delivery and collection of goods or people. Because conditions vary from one setting to the next, the objectives and constraints encountered in practice are highly variable. Most algorithmic research and software development in this area focus on a limited number of prototype problems. By building enough flexibility in optimization systems one can adapt these to various practical contexts.

Much progress has been made since the publication of the first article on the "truck dispatching" problem by Dantzig and Ramser (1959). Several variants of the basic problem have been put forward. Strong formulations have been proposed, together with polyhedral studies and exact decomposition algorithms. Numerous heuristics have also been developed for vehicle routing problems. In particular the study of this class of problems has stimulated the emergence and the growth of several metaheuristics whose performance is constantly improving.

This chapter focuses on some of the most important vehicle routing problem types. A number of other variants have been treated in recent articles and

book chapters (see, e.g., Toth and Vigo, 2002a). The *pickup and delivery vehicle routing problem*, which has also been extensively studied, is covered in the "Transportation on Demand" chapter.

The remainder of this chapter is organized as follows. Section 2 is devoted to the *classical* vehicle routing problem (simply referred to as VRP), defined with a single depot and only capacity and route length constraints. Problems with time windows are surveyed in Section 3. Section 4 is devoted to inventory routing problems which combine routing and customer replenishment decisions. Finally, Section 5 covers the field of stochastic vehicle routing in which some of the problem data are random variables.

2 The classical vehicle routing problem

The Classical Vehicle Routing Problem (VRP) is one of the most popular problems in combinatorial optimization, and its study has given rise to several exact and heuristic solution techniques of general applicability. It generalizes the Traveling Salesman Problem (TSP) and is therefore NP-hard. A recent survey of the VRP can be found in the first six chapters of the book edited by Toth and Vigo (2002a). The aim of this section is to provide a comprehensive overview of the available exact and heuristic algorithms for the VRP, most of which have also been adapted to solve other variants, as will be shown in the remaining sections.

The VRP is often defined under capacity and route length restrictions. When only capacity constraints are present the problem is denoted as CVRP. Most exact algorithms have been developed with capacity constraints in mind but several apply mutatis mutandis to distance constrained problems. In contrast, most heuristics explicitly consider both types of constraint.

2.1 Formulations

The symmetric VRP is defined on a complete undirected graph $G = (V, E)$. The set $V = \{0, \ldots, n\}$ is a vertex set. Each vertex $i \in V\backslash\{0\}$ represents a customer having a nonnegative demand q_i, while vertex 0 corresponds to a depot. To each edge $e \in E = \{(i, j)\colon\ i, j \in V, i < j\}$ is associated a travel cost c_e or c_{ij}. A fixed fleet of m identical vehicles, each of capacity Q, is available at the depot. The symmetric VRP calls for the determination of a set of m routes whose total travel cost is minimized and such that: (1) each customer is visited exactly once by one route, (2) each route starts and ends at the depot, (3) the total demand of the customers served by a route does not exceed the vehicle capacity Q, and (4) the length of each route does not exceed a preset limit L. (It is common to assume constant speed so that distances, travel times and travel costs are considered as synonymous.) A solution can be viewed as a set of m cycles sharing a common vertex at the depot. The asymmetric VRP is similarly defined on a directed graph $G = (V, A)$, where $A = \{(i, j)\colon$

$i, j \in V, i \neq j\}$ is an arc set. In this case a circuit (directed cycle) is associated with a vehicle route. Most results of Sections 2.1 and 2.2 apply to the symmetric CVRP.

An integer linear programming formulation of the CVRP follows, where for each edge $e \in E$ the integer variable x_e indicates the number of times edge e is traversed in the solution. Let $r(S)$ denote the minimum number of vehicles needed to serve the customers of a subset S of customers. The value of $r(S)$ may be determined by solving an associated Bin Packing Problem (BPP) with item set S and bins of capacity Q. Finally, for $S \subset V$, let $\delta(S) = \{(i, j)$: $i \in S, j \notin S$ or $i \notin S, j \in S\}$. If $S = \{i\}$, then we simply write $\delta(i)$ instead of $\delta(\{i\})$. The CVRP formulation proposed by Laporte et al. (1985) is then:

$$\text{(CVRP1)} \quad \text{minimize} \quad \sum_{e \in E} c_e x_e \tag{1}$$

subject to

$$\sum_{e \in \delta(i)} x_e = 2, \quad i \in V \setminus \{0\}, \tag{2}$$

$$\sum_{e \in \delta(0)} x_e = 2m, \tag{3}$$

$$\sum_{e \in \delta(S)} x_e \geqslant 2r(S), \quad S \subseteq V \setminus \{0\}, S \neq \varnothing, \tag{4}$$

$$x_e \in \{0, 1\}, \quad e \notin \delta(0), \tag{5}$$

$$x_e \in \{0, 1, 2\}, \quad e \in \delta(0). \tag{6}$$

The degree constraints (2) state that each customer is visited exactly once, whereas the depot degree constraint (3) means that m routes are created. Capacity constraints (4) impose both the connectivity of the solution and the vehicle capacity requirements by forcing a sufficient number of edges to enter each subset of vertices. We note that since the BPP is NP-hard in the strong sense, $r(S)$ may be approximated from below by any BPP lower bound, such as $\lceil \sum_{i \in S} q_i / Q \rceil$. Finally, constraints (5) and (6) impose that each edge between two customers is traversed at most once and each edge incident to the depot is traversed at most twice. In this latter case, the vehicle performs a route visiting a single customer.

A widely used alternative formulation is based on the set partitioning or set covering models. The formulation was originally proposed by Balinski and Quandt (1964) and contains a potentially exponential number of binary variables. Let $\mathcal{R} = \{R_1, \ldots, R_s\}$ denote the collection of all feasible routes, with $s = |\mathcal{R}|$. Each route R_j has an associated cost γ_j, and a_{ij} is a binary coefficient equal to 1 if and only if vertex i is visited (i.e., *covered*) by route R_j. The binary variable x_j, $j = 1, \ldots, s$, is equal to 1 if and only if route R_j is selected in the

solution. The model is:

$$\text{(CVRP2)} \quad \text{minimize} \quad \sum_{j=1}^{s} \gamma_j x_j \tag{7}$$

subject to

$$\sum_{j=1}^{s} a_{ij} x_j = 1, \quad i \in V \setminus \{0\}, \tag{8}$$

$$\sum_{j=1}^{s} x_j = m, \tag{9}$$

$$x_j \in \{0, 1\}, \quad j = 1, \ldots, s. \tag{10}$$

Constraints (8) impose that each customer i is covered by exactly one route, and (9) requires that m routes be selected. Because route feasibility is implicitly considered in the definition of $\mathcal{R}$, this is a very general model which may easily take additional constraints into account. Moreover, when the cost matrix satisfies the triangle inequality (i.e., $c_{ij} \leqslant c_{ik} + c_{kj}$ for all $i, j, k \in V$), the set partitioning model CVRP2 may be transformed into an equivalent set covering model CVRP2′ by replacing the equality sign with "$\geqslant$" in (8). Any feasible solution to CVRP2 is clearly feasible for CVRP2′, and any feasible solution to CVRP2′ may be transformed into a feasible CVRP2 solution of smaller or equal cost. Indeed, if the CVRP2′ solution is infeasible for CVRP2, then one or more customers are visited more than once. These customers may therefore be removed from their route by applying shortcuts which will not increase the solution cost because of the triangle inequality. The main advantage of using CVRP2′ is that only inclusion-maximal feasible routes, among those with the same cost, need be considered in the definition of $\mathcal{R}$. This significantly reduces the number of variables. In addition, when using CVRP2′ the dual solution space is considerably reduced since dual variables are restricted to be nonnegative. One of the main drawbacks of models CVRP2 and CVRP2′ lies in their very large number of variables, which in loosely constrained medium size instances may easily run into the billions. Thus, one has to resort to a column generation algorithm to solve these problems. The linear programming relaxation of these models tends to be very tight, as shown by Bramel and Simchi-Levi (1997). Further details on these formulations and their extensions, as well as additional formulations for the symmetric and asymmetric cases, can be found in Laporte and Nobert (1987) and in Toth and Vigo (2002b, 2002d).

2.2 Exact algorithms for the CVRP

We now review the main exact approaches presented in the last two decades for the solution of the CVRP. For a thorough review of previous exact methods, see Laporte and Nobert (1987). We first describe the algorithms based on

branch-and-bound, including those that make use of the set partitioning formulation and column generation schemes, and we then examine the algorithms based on branch-and-cut. In practice, the CVRP turns out to be significantly harder to solve than the TSP. The best CVRP algorithms can rarely tackle instances involving more than 100 vertices, while TSP instances with hundreds and even thousands of vertices are now routinely solved to optimality.

2.2.1 Branch-and-bound and set partitioning based algorithms

Several branch-and-bound algorithms are available for the solution of the CVRP. Until the late 1980s, the most effective exact methods were mainly branch-and-bound algorithms based on elementary combinatorial relaxations. Recently, more sophisticated bounds have been proposed, namely those based on Lagrangian relaxations or on the additive bounding procedure, which have substantially increased the size of the problems that can be solved to optimality. We now describe some branch-and-bound algorithms with an emphasis on lower bound computations which constitute the most critical component of methods of this type. More details on the structure of branch-and-bound algorithm strategies and dominance rules may be found in Toth and Vigo (1998, 2002c, 2002d). We also review in this section exact set partitioning based algorithms for the CVRP.

Many different elementary combinatorial relaxations were used in early branch-and-bound algorithms, including those based on the Assignment Problem (AP), on the degree-constrained *shortest spanning tree*, and on state-space relaxation. Here we outline the two families of relaxations used as a basis for the more recent branch-and-bound algorithms for the symmetric and asymmetric CVRP. A first relaxation is obtained from the integer programming formulations of these problems by dropping the connectivity and capacity constraints. In the symmetric case the resulting problem is a b-Matching Problem (b-MP), i.e., the problem of determining a minimum cost set of cycles covering all vertices and such that the degree of each vertex i is equal to b_i, where $b_i = 2$ for all the customer vertices, and $b_0 = 2m$ for the depot vertex. It is easy to see that by adding $m - 1$ copies of the depot to G the relaxation becomes a 2-MP. In the asymmetric case the relaxed problem is the well-known transportation problem which may be transformed into an AP by introducing copies of the depot. Also in this case, the AP may be seen as the problem of determining a set of circuits covering all vertices and such that each vertex has one entering and one leaving arc. The solution of these relaxed problems may be infeasible for the CVRP since the demand associated with a cycle or circuit may exceed the vehicle capacity, and some of these may be disconnected from the depot. The relaxed problems may then be solved in polynomial time (see, e.g., Miller and Pekny, 1995, for the b-MP and Dell'Amico and Toth, 2000 for the AP). However, the quality of the lower bounds obtained with these relaxations is generally very poor and not sufficient to solve instances with more than 15 or 20 customers. Toth and Vigo (2002c) report average gaps in excess of 20% with respect to the optimal solution value on benchmark CVRP in-

stances. The situation is slightly better for the AP relaxation of the asymmetric CVRP that yields average gaps of about 10% or less. Laporte et al. (1986) have proposed a branch-and-bound algorithm for asymmetric CVRP, based on the AP relaxation and capable of solving randomly generated problems involving tens of customers and between two and four vehicles.

The second family of elementary relaxations used in recent branch-and-bound algorithms is based on degree-constrained spanning trees. These relaxations extend the well-known 1-tree relaxation proposed by Held and Karp (1971) for the TSP. The earliest branch-and-bound algorithm based on this relaxation, proposed by Christofides et al. (1981a), could only solve relatively small instances. More recently, Fisher (1994) has presented another tree based relaxation requiring the determination of a so-called m-tree, defined as a minimum cost set of $n + m$ edges spanning the graph. The approach used by Fisher is based on CVRP1 with the additional assumption that single-customer routes are not allowed. Fisher modeled the CVRP as the problem of determining an m-tree with degree equal to $2m$ at the depot vertex, with additional constraints on vehicle capacity and a degree of 2 for each customer vertex. The determination of an m-tree with degree $2m$ at the depot requires $O(n^3)$ time. The degree-constrained m-tree relaxation is easily obtained from CVRP1 by removing the degree constraints (2) for customer vertices and weakening the capacity constraints (4) into connectivity constraints, i.e., by replacing their right-hand side with 1. The m-tree solution is not always feasible for the CVRP since some vertices may have a degree different from 2 and the demand associated with the subtrees incident to the depot may exceed the vehicle capacity.

For the asymmetric CVRP, similar relaxations may be derived from directed trees, also called *arborescences*, spanning the graph and having an outdegree equal to m at the depot vertex. To obtain the final bound a minimum cost set of m vertex-disjoint arcs entering the depot are added to the constrained arborescence. In this case, the relaxed subproblem may be solved in polynomial time, but again the quality of the resulting lower bound is very poor. Toth and Vigo (2002c) report that on benchmark asymmetric instances, the average gap of these relaxations with respect to the optimal solution value is larger than 25%.

Different improved bounding techniques were later developed to narrow the gap between the lower bound and the optimal solution value of the CVRP. These include two bounding procedures based on Lagrangian relaxation proposed by Fisher (1994) and Miller (1995). These are strengthenings of the basic CVRP relaxations obtained by dualizing some of the relaxed constraints in a Lagrangian fashion. In particular, they both include in the objective function a suitable subset of the capacity constraints (4), whereas the Fisher relaxation also incorporates degree constraints (2) which were relaxed in the m-tree relaxation. As in related problems, good values for the Lagrangian multipliers associated with the relaxed constraints are determined by using a subgradient optimization procedure (see, e.g., Held and Karp, 1971; Held et al., 1974). The main difficulty associated with these relaxations lies in the exponential cardinality of the set of relaxed constraints which does not

allow for their complete inclusion in the objective function. These authors include a limited family $\mathcal{F}$ of capacity constraints and iteratively generate the constraints violated by the current solution of the Lagrangian problem. The process terminates when no violated constraint is detected (hence the Lagrangian solution is feasible) or a preset number of subgradient iterations have been executed. Redundant constraints are periodically removed from $\mathcal{F}$. The relax-and-cut algorithm of Martinhon et al. (2000) generalizes these Lagrangian-based approaches by also considering comb and multistar inequalities, and moderately improves the quality of the Lagrangian bound.

Some exact algorithms for the CVRP are based on the set partitioning formulation CVRP2. The first of these is due to Agarwal et al. (1989) who considered a relaxation of model CVRP2 not including constraints (9) and solved the resulting model through column generation. Agarwal, Mathur, and Salkin used their algorithm to solve seven Euclidean CVRP instances with up to 25 customers. Hadjiconstantinou et al. (1995) proposed a branch-and-bound algorithm in which the lower bound was obtained by considering the dual of the linear relaxation of model CVRP2, following the approach introduced by Mingozzi et al. (1994). By linear programming duality, any feasible solution to this dual problem yields a valid lower bound. Hadjiconstantinou et al. (1995) determined the heuristic dual solutions by combining two relaxations of the original problem: the q-path relaxation of Christofides et al. (1981a) and the m-shortest path relaxation of Christofides and Mingozzi (1989). The algorithm was able to solve randomly generated Euclidean instances with up to 30 vertices and benchmark instances with up to 50 vertices. Further details on set partitioning-based algorithms for the CVRP are provided in Bramel and Simchi-Levi (2002).

Fischetti et al. (1994) have improved the AP relaxation of the asymmetric CVRP by combining into an additive bounding procedure two new lower bounds based on disjunctions on infeasible arc subsets and on minimum cost flows. The additive approach was proposed by Fischetti and Toth (1989) and allows for the combination of different lower bounding procedures, each exploiting a different substructure of the problem under consideration. The resulting branch-and-bound approach was able to solve randomly generated instances containing up to 300 vertices and four vehicles. Other bounds for the asymmetric CVRP may be derived by generalizing the methods proposed for the symmetric case. For example, Fisher (1994) proposed a way of extending to the asymmetric CVRP the Lagrangian bound based on m-trees. In this extension the Lagrangian problem calls for the determination of an undirected m-tree on the undirected graph obtained by replacing each pair of arcs (i, j) and (j, i) with a single edge (i, j) of cost $c'_{ij} = \min\{c_{ij}, c_{ji}\}$. No computational testing for this bound was presented by Fisher (1994). Potentially better bounds may be obtained by explicitly considering the asymmetry of the problem, i.e., by using m-arborescences rather than m-trees and by strengthening the bound in a Lagrangian fashion as proposed by Toth and Vigo (1995, 1997) for the *capacitated shortest spanning arborescence problem* and for the *VRP with backhauls*.

2.2.2 *Branch-and-cut algorithms*

Branch-and-cut algorithms currently constitute the best available exact approach for the solution of the CVRP. Research in this area has been strongly motivated by the emergence and the success of polyhedral combinatorics as a framework for the solution of hard combinatorial problems, particularly the TSP. However, in a recent survey on branch-and-cut approaches for the CVRP, Naddef and Rinaldi (2002) state: "... the amount of research effort spent to solve CVRP by this method is not comparable with what has been dedicated to the TSP [... the research in this field] is still quite limited and most of it is not published yet". In the following we summarize the main available branch-and-cut approaches for the CVRP. The reader is referred to Naddef and Rinaldi (2002) for a more detailed presentation.

The use of branch-and-cut for the CVRP is rooted in the exact algorithm of Laporte et al. (1985). This algorithm uses the Linear Programming (LP) relaxation of model CVRP1 without capacity constraints (4) as a basis for the solution of the VRP with capacity and maximum distance restrictions. This initial relaxation is iteratively strengthened by adding violated capacity constraints which are heuristically separated by considering the connected components induced by the set of nonzero variables in the current LP solution. Gomory cuts are also introduced at the root node of the branch-and-cut tree. The algorithm was capable of solving randomly generated loosely constrained Euclidean and non-Euclidean instances with two or three vehicles and up to 60 customers.

The first polyhedral study of the CVRP was presented by Cornuéjols and Harche (1993). The presence of equalities (2) and (3) makes the CVRP nonfully-dimensional. Therefore, as in the TSP, Cornuéjols and Harche first considered the full-dimensional polyhedron, containing the CVRP polyhedron as a face, associated with the so-called Graphical VRP (GVRP) where customers may be visited more than once. The basic properties of the GVRP polyhedron were also investigated. Conditions under which the nonnegativity, degree and capacity constraints define facets of the GVRP and CVRP polyhedra were also determined. Cornuéjols and Harche have extended to the GVRP and the CVRP several other families of valid inequalities proposed for the TSP and the graphical TSP. In particular, comb, path, wheelbarrow, and bicycle inequalities were extended to the capacitated case and again, sufficient conditions under which these inequalities define facets of the GVRP and CVRP polyhedra were derived. These inequalities were used by Cornuéjols and Harche as cutting planes to solve two instances of CVRP with 18 and 50 customers, within a branch-and-cut algorithm. The detection of violated inequalities was performed manually, starting from the current optimal LP solution.

Augerat et al. (1995) have developed the first complete branch-and-cut approach for the CVRP. They described several heuristic separation procedures for the classes of valid inequalities proposed by Cornuéjols and Harche, as well as four new classes of valid inequalities. Separation procedures were further investigated by Augerat et al. (1999). The resulting approach was able to solve

several CVRP instances containing up to 134 customers. Ralphs et al. (2003) have presented a branch-and-cut algorithm for the CVRP in which an exact separation of valid m-TSP inequalities is used in addition to heuristic separation of capacity inequalities. The resulting algorithm was implemented within the SYMPHONY parallel branch-and-cut-and-price framework and was able to solve several instances involving fewer than 100 vertices. Lysgaard et al. (2004) have developed new separation procedures for most of the families of valid inequalities proposed so far (see also Letchford et al., 2002). Their overall branch-and-cut approach, which is further enhanced by the use of Gomory cuts, was able to solve within moderate computing times previously solved instances and three new medium size ones.

Baldacci et al. (2004) have put forward a branch-and-cut algorithm based on a two-commodity network flow formulation of the CVRP and requiring a polynomial number of integer variables. It seems to provide an interesting alternative to other classical formulations (see also Gouveia, 1995, for a single-commodity formulation). The overall algorithm strengthens the LP relaxation by adding violated capacity inequalities and implements various variable reduction and branching rules. The results obtained with this approach are comparable with those of the other branch-and-cut algorithms just described.

Finally, Fukasawa et al. (2006) have proposed a successful branch-and-cut-and-price algorithm combining branch-and-cut with the q-routes relaxation of Christofides et al. (1981a), used here in a column generation fashion. This method produces tighter bounds than other branch-and-cut algorithms and is capable of solving several previously unsolved instances with up to 75 customers. Baldacci et al. (2006) have used their set partitioning algorithm, previously developed for a rollon–rolloff VRP, to solve difficult CVRP instances. Their approach yields bounds whose quality is comparable to those of Fukasawa et al. (2006), but seems much quicker.

Other branch-and-cut algorithms are described in Achuthan et al. (1996, 2003) and Blasum and Hochstättler (2000). We also mention that the polyhedral structure of the special case of CVRP where all the customers have a unit demand was studied by Campos et al. (1991) and by Araque et al. (1990). Branch-and-cut algorithms for this problem are presented by Araque et al. (1994) and by Ghiani et al. (2006).

2.3 Heuristics for the VRP

An impressive number of heuristics have been proposed for the VRP. Initially these were mainly standard route construction algorithms, whereas more recently powerful metaheuristic approaches have been developed. In the following we separately review these two families of algorithms. Almost all of these methods were developed, described and tested for the symmetric VRP. In addition, since finding a feasible solution with exactly m vehicles is itself an NP-complete problem, almost all methods assume an unlimited number

of available vehicles. However, it should be observed that many of the proposed methods may be quite easily adapted to take into account additional practical constraints, although these may affect their overall performance (see, e.g., Vigo, 1996, for an extension of some classical heuristics to the asymmetric case).

2.3.1 *Classical heuristics*

Using the classification proposed by Laporte and Semet (2002), we describe classical VRP heuristics under these headings: route construction methods, two-phase methods, and route improvement methods.

Route construction heuristics. Route construction methods were among the first heuristics for the CVRP and still form the core of many software implementations for various routing applications. These algorithms typically start from an empty solution and iteratively build routes by inserting one or more customers at each iteration, until all customers are routed. Construction algorithms are further subdivided into sequential and parallel, depending on the number of eligible routes for the insertion of a customer. Sequential methods expand only one route at a time, whereas parallel methods consider more than one route simultaneously. Route construction algorithms are fully specified by their three main ingredients, namely an initialization criterion, a selection criterion specifying which customers are chosen for insertion at the current iteration, and an insertion criterion to decide where to locate the chosen customers into the current routes.

The first and most famous heuristic of this group was proposed by Clarke and Wright (1964) and is based on the concept of *saving*, an estimate of the cost reduction obtained by serving two customers sequentially in the same route, rather than in two separate ones. If i is the last customer of a route and j is the first customer of another route, the associated saving is defined as $s_{ij} = c_{i0} + c_{0j} - c_{ij}$. If s_{ij} is positive, then serving i and j consecutively in a route is profitable. The Clarke and Wright algorithm considers all customer pairs and sorts the savings in nonincreasing order. Starting with a solution in which each customer appears separately in a route, the customer pair list is examined and two routes are merged whenever this is feasible. Generally, a route merge is accepted only if the associated saving is nonnegative but, if the number of vehicles is to be minimized, then negative saving merges may also be considered. The Clarke and Wright algorithm is inherently parallel since more than one route is active at any time. However, it may easily be implemented in a sequential fashion. The resulting algorithm is quite fast but may have a poor performance (see, e.g., Laporte and Semet, 2002). Golden et al. (1977), Paessens (1988), and Nelson et al. (1985) have proposed various enhancement strategies of the savings approach aimed at improving either its effectiveness or its computational efficiency by means of better data structures. Other attempts to improve the effectiveness of the savings method were made by Desrochers and Verhoog (1989), Altinkemer and Gavish (1991), and by Wark and Holt

(1994) who proposed to implement route merges by using a matching algorithm, together with a more sophisticated estimate of actual merge savings. The results obtained with these algorithms are in general better than those of previous savings methods, but matching-based algorithms require much larger computing times.

Another classical route construction heuristic is the sequential insertion algorithm of Mole and Jameson (1976). The algorithm uses as selection and insertion criterion the evaluation of the extra distance resulting from the insertion of an unrouted customer k between two consecutive customers i and j of the current route, namely $\alpha(i, k, j) = c_{ik} + c_{kj} - \lambda c_{ij}$, where λ is a user-controlled parameter. Variations of this criterion taking into account other factors, such as the distance of the customer from the depot, were also considered. After each insertion, the current route is possibly improved by using a 3-opt procedure. A more general and effective two-step insertion heuristic was proposed by Christofides et al. (1979). In the first step, a sequential insertion algorithm is used to determine a set of feasible routes. The second step is a parallel insertion approach. For each route determined in the first step, a representative customer is selected and a set of single-customer routes is initialized with these customers. The remaining unrouted customers are then inserted by using a regret criterion, where the difference between the best and the second-best insertion cost is taken into account, and partial routes are improved by means of a 3-opt procedure. The resulting algorithm is superior to that of Mole and Jameson and represents a good compromise between effectiveness and efficiency.

Two-phase heuristics. Two-phase methods are based on the decomposition of the VRP solution process into the two separate subproblems:

(1) clustering: determine a partition of the customers into subsets, each corresponding to a route, and
(2) routing: determine the sequence of customers on each route.

In a cluster-first-route-second method, customers are first grouped into clusters and the routes are then determined by suitably sequencing the customers within each cluster. Different techniques have been proposed for the clustering phase, while the routing phase amounts to solving a TSP.

The *sweep* algorithm, due to Wren (1971), Wren and Holliday (1972), and Gillett and Miller (1974), is often referred to as the first example of cluster-first-route-second approach. The algorithm applies to planar VRP instances. The algorithm starts with an arbitrary customer and then sequentially assigns the remaining customers to the current vehicle by considering them in order of increasing polar angle with respect to the depot and the initial customer. As soon as the current customer cannot be feasibly assigned to the current vehicle, a new route is initialized with it. Once all customers are assigned to vehicles, each route is separately defined by solving a TSP. Another early two-phase method is the truncated branch-and-bound method of Christofides et al.

(1979) in which the set of routes is determined through an adaptation of an exact branch-and-bound algorithm that uses a branching-on-routes strategy. The decision tree contains as many levels as the number of available vehicles, and at each level of the decision tree a given node corresponds to a partial solution made up of some complete routes. The descendant nodes correspond to all possible routes including a subset of the unrouted customers. The running time of the algorithm is controlled by limiting to one the number of routes generated at each level.

The Fisher and Jaikumar (1981) algorithm solves the clustering step by means of an appropriately defined Generalized Assignment Problem (GAP) which calls for the determination of a minimum cost assignment of items to a given set of bins of capacity Q, and where the items are characterized by a weight and an assignment cost for each bin. Each vehicle is assigned a representative customer, called a *seed*, and the assignment cost of a customer to a vehicle is equal to its distance to the seed. The GAP is then solved, either optimally or heuristically, and the final routes are determined by solving a TSP on each cluster.

Another two-phase method working with a fixed number m of vehicles was described by Bramel and Simchi-Levi (1995). This algorithm determines route seeds by solving a capacitated location problem, where m customers are selected by minimizing the total distance between each customer and its closest seed, and by imposing that the total demand associated with each seed be at most Q. Once seeds have been determined and the single-customer routes are initialized, the remaining customers are inserted in the current routes by minimizing insertion costs. Various ways of approximating the insertion cost are proposed and analyzed. It is worth noting that all three cluster-first-route-second approaches just described allow for a direct control of the number of routes in the final solution, whereas the sweep algorithm does not. The performance of these algorithms is generally comparable to that of route construction algorithms in terms of effectiveness. The location based approach of Bramel and Simchi-Levi produces better solutions but requires much larger computing times.

A different family of two-phase methods is the class of so-called *petal* algorithms. These generate a large set of feasible routes, called petals, and select the final subset by solving a set partitioning model. Foster and Ryan (1976) and Ryan et al. (1993) have proposed heuristic rules for determining the set of routes to be selected, while Renaud et al. (1996b) have described an extension that considers more involved configurations, called 2-petals, consisting of two embedded or intersecting routes. The overall performance of these algorithms is generally superior to that of the sweep algorithm.

Finally, in route-first-cluster-second methods, a giant TSP tour over all customers is constructed in a first phase and later subdivided into feasible routes. Examples of such algorithms are given by Beasley (1983), Haimovich and Rinnooy Kan (1985), and Bertsimas and Simchi-Levi (1996), but the performance of this approach is generally poor.

Route improvement heuristics. Local search algorithms are often used to improve initial solutions generated by other heuristics. Starting from a given solution, a local search method applies simple modifications, such as arc exchanges or customer movements, to obtain neighbor solutions of possibly better cost. If an improving solution is found, it then becomes the current solution and the process iterates; otherwise a local minimum has been identified.

A large variety of neighborhoods are available. These may be subdivided into *intra-route* neighborhoods, if they operate on a single route at a time, or *inter-route* neighborhoods if they consider more than one route simultaneously. The most common neighborhood type is the λ-opt heuristic of Lin (1965) for the TSP, where λ edges are removed from the current solution and replaced by λ others. The computing time required to examine all neighbors of a solution is proportional to n^{λ}. Thus, only $\lambda = 2$ or 3 are used in practice. As an alternative, one can use restricted neighborhoods characterized by subsets of moves associated with larger λ values, such as Or-exchanges (Or, 1976) or the 4-opt* neighborhood of Renaud et al. (1996a) which considers only a subset of all potential 4-opt exchanges. Laporte and Semet (2002) have conducted a computational comparison of some basic route improvement procedures. More complex inter-route neighborhoods are analyzed by Thompson and Psaraftis (1993), Van Breedam (1994), and Kindervater and Savelsbergh (1997).

2.3.2 *Metaheuristics*

Several metaheuristics have been applied to the VRP. With respect to classical heuristics, they perform a more thorough search of the solution space and are less likely to end with a local optimum. These can be broadly divided into three classes:

(1) local search, including simulated annealing, deterministic annealing, and tabu search;
(2) population search, including genetic search and adaptive memory procedures;
(3) learning mechanisms, including neural networks and ant colony optimization.

The best heuristics often combine ideas borrowed from different metaheuristic principles. Recent surveys of VRP metaheuristics can be found in Gendreau et al. (2002), Cordeau and Laporte (2004), and Cordeau et al. (2005).

Local search algorithms explore the solution space by iteratively moving from a solution x_t at iteration t to a solution x_{t+1} in the neighborhood $N(x_t)$ of x_t until a stopping criterion is satisfied. If $f(x)$ denotes the cost of solution x, then $f(x_{t+1})$ is not necessarily smaller than $f(x_t)$. As a result, mechanisms must be implemented to avoid cycling. In simulated annealing, a solution x is drawn randomly from $N(x_t)$. If $f(x) \leqslant f(x_t)$, then $x_{t+1} := x$. Otherwise,

$$x_{t+1} := \begin{cases} x & \text{with probability } p_t, \\ x_t & \text{with probability } 1 - p_t, \end{cases}$$

where p_t is a decreasing function of t and of $f(x) - f(x_t)$. This probability is often equal to

$$p_t = \exp\left(-\frac{f(x) - f(x_t)}{\theta_t}\right),$$

where θ_t is the *temperature* at iteration t, usually defined as a nonincreasing function of t. Deterministic annealing (Dueck, 1990, 1993) is similar. There are two main versions of this algorithm: in a threshold-accepting algorithm, $x_{t+1} := x$ if $f(x) < f(x_t) + \theta_1$, where θ_1 is a user controlled parameter; in record-to-record travel, a record is the best known solution x^*, and $x_{t+1} := x$ if $f(x_{t+1}) < \theta_2 f(x^*)$, where θ_2 is also user controlled. In tabu search, in order to avoid cycling, any solution possessing some given attribute of x_{t+1} is declared tabu for a number of iterations. At iteration t, the search moves to the best nontabu solution x in $N(x_t)$. These local search algorithms are rarely implemented in their basic version, and their success depends on the careful implementation of several mechanisms. The rule employed to define neighborhoods is critical to most local search heuristics. In simulated annealing several rules have been proposed to define θ_t (see Osman, 1993). Tabu search relies on various strategies to implement tabu tenures (also known as short term memory), search diversification (also known as long term memory), and search intensification which accentuates the search in a promising region.

Population search algorithms operate on several generations of solution populations. In genetic search it is common to repeat the following operation k times: extract two parent solutions from the populations to create two offspring using a crossover operation, and apply a mutation operation to each offspring; then remove the $2k$ worst elements from the population and replace them with the $2k$ offspring. Several crossover rules are available for sequencing problems (Bean, 1994; Potvin, 1996; Drezner, 2003; Prins, 2004). In adaptive memory procedures, an offspring is created by extracting and recombining elements of several parents. In the initial version proposed by Rochat and Taillard (1995) for the VRP, nonoverlapping routes are extracted from several parents to create a partial solution. This solution is then gradually completed and optimized by tabu search.

Neural networks are models composed of richly interconnected units through weighted links, like neurons in the brain. They gradually construct a solution through a feedback mechanism that modifies the link weights to better match an observed output to a described output. In the field of vehicle routing neural network models called the elastic net and the self-organizing map are deformable templates that adjust themselves to the contour of the vertices to generate a feasible VRP solution. An example is provided by Ghaziri (1993). Ant colony algorithms (see Dorigo et al., 1999) also use a learning mechanism. They are derived from an analogy with ants which lay some pheromone on their trail when foraging for food. With time more pheromone is deposited on the more frequented trails. When constructing a VRP solution a move can

be assigned a higher probability of being selected if it has previously led to a better solution in previous iterations.

In what follows we summarize the most effective metaheuristics for the CVRP. Initially the best methods were almost exclusively based on tabu search but in recent years several excellent methods inspired from different paradigms have been proposed.

Local search heuristics. A limited number of simulated annealing heuristics for the CVRP were proposed in the early 1990s. Osman's implementation (Osman, 1993) is the most involved and also the most successful. It defines neighborhoods by means of a 2-interchange scheme and applies a different rule of temperature changes. Instead of using a nonincreasing function, as do most authors in the field, Osman decreases θ_t continuously as long as the solution improves, but whenever $x_{t+1} = x_t$, θ_t is either halved or replaced by the temperature at which the incumbent was identified. This algorithm succeeded in producing good solutions but was not competitive with the best tabu search implementations available at the same period.

A large number of tabu search algorithms have been produced over the past fifteen years (a survey is available in Cordeau and Laporte, 2004). In the first known implementation, due to Willard (1989), a CVRP solution is represented as a giant tour containing several copies of the depot and inter-depot chains corresponding to feasible vehicle routes, and neighborhoods are defined by means of 3-opt exchanges. The method was soon to be superseded by more powerful algorithms, including those of Osman (1993), Taillard (1993), and Gendreau et al. (1994).

Taillard's algorithm remains to this day one of the most successful tabu search implementations for the CVRP. It is based on the use of an 1-interchange mechanism to define neighbor solutions, combined with periodic route reoptimizations by means of an exact TSP algorithm (Volgenant and Jonker, 1983). The algorithm also uses random tabu durations. A continuous diversification mechanism that penalizes frequently performed moves is implemented in order to provide a more thorough exploration of the search space. Finally, Taillard's algorithm employs a decomposition scheme that allows for the use of parallel computing. In planar problems the customer set is partitioned into sectors and then concentric rings, while in random instances the regions are defined by means of shortest spanning arborescences rooted at the depot. The region boundaries are periodically updated to produce a diversification effect.

The Taburoute algorithm of Gendreau et al. (1994) moves at each iteration a vertex from its current route to another route containing one of its closest neighbors. Insertions are performed simultaneously with a local reoptimization of the route, based on the GENI procedure (Gendreau et al., 1992). Only a subset of vertices are considered for reinsertion at any given iteration. No vertex can return to its former route during the next θ iterations, where θ is randomly selected in a closed interval. Taburoute also uses

a continuous diversification mechanism. During the course of the search infeasible solutions are penalized. This mechanism is implemented by replacing the solution value $f(x)$ associated at solution x with a penalized objective $f'(x) = f(x) + \alpha Q(x) + \beta L(x)$, where $Q(x)$ is the total capacity violation of solution x and $L(x)$ is the total route length violation. The two parameters α and β self-adjust during the search to produce a mix of feasible and infeasible solutions: every μ iterations, α (resp. β) is divided by 2 if the past μ solutions were feasible with respect to capacity (resp. route length), or multiplied by 2 if they were all infeasible with respect to capacity (resp. route length). Other features of Taburoute include the use of random tabu durations, periodic route reoptimizations by means of the US procedure of Gendreau et al. (1992), false starts to initialize the search, and a final intensification phase around the best known solution.

The Rego and Roucairol (1996) Tabuchain algorithm is based on the use of ejection chains involving ℓ routes to define neighborhoods. This process bumps a vertex from one route of the chain to another route. The last bumped vertex may be relocated in the position of the first bumped vertex or elsewhere. The process ensures that no arc or edge is considered more than once in the solution. As in Taburoute, intermediate infeasible solutions are allowed. The authors have also implemented a sequential and a parallel version of their method. Another ejection scheme, called Flower, was later developed by Rego (1998). It is based on the idea of exploiting the representation of routes as blossoms and of paths as stems, and of performing ejection moves by means of edge deletions and creations. This method was not as successful as Tabuchain. Another method employing ejection chains was developed by Xu and Kelly (1996). It oscillates between ejection chains and vertex swaps between two routes. The ejection chains are obtained by solving an auxiliary network flow problem. On the whole this method succeeded in obtaining several good CVRP solutions on benchmark instances but it is rather involved and time consuming.

More recently, Ergun et al. (2003) have developed a Very Large Neighborhood Search (VLNS) algorithm for the VRP. This algorithm operates on several routes simultaneously, not unlike what is done in cyclic transfers (Thompson and Psaraftis, 1993) or in ejection chains. Neighborhoods are defined by a combination of 2-opt moves, vertex swaps between routes, and vertex insertions in different routes. The best choice of moves and of routes involved in the moves is determined through the solution of a network flow problem on an auxiliary graph. One advantage of VLNS is that it allows a broad search by acting on several routes at once. Its main disadvantage lies in the effort required at each iteration to perform moves.

A very useful concept put forward by Toth and Vigo (2003) is that of Granular Tabu Search (GTS). This algorithm a priori removes from the graph long edges that are unlikely to belong to an optimal solution. To determine these edges, the problem is first solved by means of a fast heuristic, e.g., the Clarke and Wright (1964) algorithm, and the average edge cost $\bar{c}$ in this solution

is determined. Then only two families of edges are retained: those incident to the depot, and those whose cost does not exceed $\beta\bar{c}$, where β is a user-defined sparsification parameter. The authors show that on benchmark instances, choosing β in [1.0, 2.0] yields the elimination of between 80–90% of all edges. Granular tabu search was implemented in conjunction with some of the features of Taillard's algorithm (Taillard, 1993) and Taburoute (Gendreau et al., 1994), and neighbor solutions were obtained by performing intra-route and inter-route exchanges.

Deterministic annealing was first applied to the VRP by Golden et al. (1998) and more recently by Li et al. (2005). The latter algorithm combines the record-to-record principle of Dueck (1993) with GTS. It works on a sparsified graph containing only a proportion α of the 40 shortest edges incident to each vertex, where α varies throughout the algorithm. The algorithm is applied several times from three initial solutions generated by the Clarke and Wright (1964) algorithm, with savings s_{ij} defined as $c_{i0} + c_{0j} - \lambda c_{ij}$, and $\lambda = 0.6, 1.4,$ and 1.6. Neighbors are defined by means of intra- and inter-route 2-opt moves, and nonimproving solutions are accepted as long as their cost does not exceed that of the incumbent by more than 1%. Whenever the solution has not improved for a number of iterations, a perturbation is applied to the best known solution to restart the search. This is achieved by temporarily moving some vertices to different positions.

Population search heuristics. The Adaptive Memory Procedure (AMP) put forward by Rochat and Taillard (1995) constitutes a major contribution to the field of metaheuristics. Initially developed in the context of the VRP, it is of general applicability and has been used, for example, to solve political districting problems (Bozkaya et al., 2003). An adaptive memory is a pool of good solutions which is updated by replacing its worst elements with better ones. In order to generate a new solution, several solutions are selected from the pool and recombined. In the context of the VRP, vehicle routes are extracted from these solutions and used as the basis of a new solution. The extraction process is applied as long as it is possible to identify routes that do not overlap with previously selected routes. When this is no longer possible, a search process (e.g., tabu search) is initiated from a partially constructed solution made up of the selected routes and some unrouted customers. Any solution constructed in this fashion replaces the worst solution of the pool if it has a better cost. Tarantilis and Kiranoudis (2002) have proposed a variant to this scheme. In a first phase a solution is obtained by means of the Paessens (1988) constructive procedure, which is an application of the Clarke and Wright savings heuristic followed by 2-opt moves, vertex swaps between routes, and vertex reinsertions. In order to generate new solutions from the adaptive memory, Tarantilis and Kiranoudis extract route segments, called bones, as opposed to full vehicle routes as did Rochat and Taillard.

Prins (2004) has developed an algorithm combining two main features of evolutionary search, namely crossovers and mutations. Crossovers consist of

creating offspring solutions from parents, while mutations are obtained here by applying a local search algorithm to an offspring. This combination of solution recombination and local search is sometimes referred to as a memetic algorithm (Moscato and Cotta, 2003). In this algorithm, solutions are represented as a giant tour without trip delimiters. To create an offspring from two parents, a chain $(i, \ldots, j)$ is first selected from the first parent and the vertices of the second parent are scanned from position $j + 1$ by skipping those of the chain $(i, \ldots, j)$. A second offspring is generated in a similar way by reversing the roles of the two parents. Offspring are improved by applying a combination of vertex and edge reinsertions, vertex swaps, combined vertex and edge swaps.

Two other memetic algorithms have recently been proposed by Berger and Barkaoui (2004) and by Mester and Bräysy (2005). The first works on two populations whose sizes are kept constant through the replacement of parents by newly created offspring, and migrations take place between the two populations. Offspring are obtained by combining routes from two parents as long as this can be done without overlapping, and by inserting the unrouted customers according to a proximity criterion. A VLNS heuristic (Shaw, 1998) combining three insertion mechanisms is then applied to the offspring, followed by an improvement scheme consisting of removing vertices from the solution and reinserting them by means of the I1 procedure of Solomon (1987).

The Active Guided Evolution Strategies (AGES) of Mester and Bräysy was initially developed to solve the VRP with time windows and was later applied to the classical VRP. It combines local search (Voudouris, 1997) with an evolution strategy (Rechenberg, 1973) to produce an iterative two-stage procedure. The evolutionary strategy uses a deterministic rule to select a parent solution and create a single offspring from a single parent. The offspring replaces the parent if it improves upon it. Offspring are improved by means of an elaborate search procedure combining granular tabu search, continuous diversification, vertex swaps and moves, 2-opt* moves (Potvin and Rousseau, 1995), VLNS (Shaw, 1998), and restarts.

Learning mechanisms. A limited number of heuristics based on learning mechanisms have been proposed for the VRP. None of the known neural networks based methods is satisfactory, and the early ant colony based heuristics could not compete with the best available approaches. Recently, however, Reimann et al. (2004) have proposed a well-performing heuristics called D-ants. The method repeatedly applies two phases until a stopping criterion is reached. In the first phase, a first generation of good solutions is generated through the applications of a savings based heuristic (Clarke and Wright, 1964) and a 2-opt improvement procedure is applied to each solution. New generations of solutions are then created by benefiting from the knowledge gained in producing past generations. Thus, instead of using the standard savings $s_{ij} = c_{i0} + c_{0j} - c_{ij}$, an attractiveness value $\chi_{ij} = \tau_{ij}^{\alpha} s_{ij}^{\beta}$ is now employed, where τ_{ij}^{α} contains information on how good linking i and j turned out to be in previous generations, and α and β are user-controlled parameters. Vertices

i and j are linked with probability $p_{ij} = \chi_{ij}/(\sum_{(h,\ell)\in\Omega_k} \chi_{h\ell})$, where Ω_k is the set of the feasible (i, j) pairs yielding the k best savings. In the second phase the best solution identified in the first phase is decomposed into subproblems which are then reoptimized using the procedure used in the first phase.

Computational comparison of metaheuristics. Cordeau et al. (2005) provide a computational comparison of recent VRP heuristics on the 14 Christofides et al. (1979) instances ($50 \leqslant n \leqslant 199$) and on the 20 larger Li et al. (2005) instances ($200 \leqslant n \leqslant 480$). Most metaheuristics used in the comparison consistently yield solutions whose value lies within 1% of the best known value.

On the Christofides et al. (1979) instances, the best solutions are obtained by Taillard (1993), Rochat and Taillard (1995), and Mester and Bräysy (2005). If the two instance sets are taken together, the best performers, in terms of accuracy and computing time are probably Mester and Bräysy (2005), Tarantilis and Kiranoudis (2002), and Prins (2004). It should be noted that these three methods all combine population search and local search, thus allowing for a broad and deep exploration of the solution space.

As noted by Cordeau et al. (2002b) heuristics should not be judged solely on speed and accuracy. Simplicity and flexibility are also important. In this respect the Li et al. (2005) record-to-record algorithm is rather interesting: this algorithm possesses a simple structure and is capable of generating very high quality solutions. As far as flexibility is concerned, the granularity principle (Toth and Vigo, 2003) and the adaptive memory concept (Rochat and Taillard, 1995) are general and useful ideas which can easily be applied to other problems.

3 The vehicle routing problem with time windows

The Vehicle Routing Problem with Time Windows (VRPTW) is an important generalization of the classical VRP in which service at every customer i must start within a given time window $[a_i, b_i]$. A vehicle is allowed to arrive before a_i and wait until the customer becomes available, but arrivals after b_i are prohibited. The VRPTW has numerous applications in distribution management. Common examples are beverage and food delivery, newspaper delivery, and commercial and industrial waste collection (see, e.g., Golden et al., 2002).

The VRPTW is NP-hard since it generalizes the CVRP which is obtained when $a_i = 0$ and $b_i = \infty$ for every customer i. In the case of a fixed fleet size, even finding a feasible solution to the VRPTW is itself an NP-complete problem (Savelsbergh, 1985). As a result, research on the VRPTW has concentrated on heuristics. Nevertheless, when the problem is sufficiently constrained (i.e., when time windows are sufficiently narrow), realistic size instances can be solved optimally through mathematical programming techniques. This section presents a mathematical formulation of the VRPTW followed by a description of some of the most important available exact and heuristic algorithms. It is

worth pointing out that while exact methods usually minimize distance, most heuristics consider a hierarchical objective which first minimizes the number of vehicles used and then distance.

3.1 Formulation of the VRPTW

The VRPTW can be defined on a directed graph $G = (V, A)$, where $|V| = n + 2$, and the depot is represented by the two vertices 0 and $n + 1$. Feasible vehicle routes then correspond to paths starting at vertex 0 and ending at vertex $n + 1$. The set of vehicles is denoted by K, with $|K| = m$. Let s_i denote the service time at i (with $s_0 = s_{n+1} = 0$) and let t_{ij} be the travel time from i to j. In addition to the time window $[a_i, b_i]$ associated with each vertex $i \in N = V \setminus \{0, n+1\}$, time windows $[a_0, b_0]$ and $[a_{n+1}, b_{n+1}]$ can also be associated with the depot vertex. If no particular restrictions are imposed on vehicle availability, one may simply set $a_0 = \min_{i\in N}\{a_i - t_{0i}\}$, $b_0 = \max_{i\in N}\{b_i - t_{0i}\}$, $a_{n+1} = \min_{i\in N}\{a_i + s_i + t_{i,n+1}\}$, and $b_{n+1} = \max_{i\in N}\{b_i + s_i + t_{i,n+1}\}$. As in the CVRP, let q_i denote the demand of customer i, and let Q be the vehicle capacity.

While several models are available for the VRPTW, this problem is often formulated as a multicommodity network flow model with time window and capacity constraints. This model involves two types of variables: binary variables x_{ij}^k, $(i, j) \in A$, $k \in K$, equal to 1 if and only if arc (i, j) is used by vehicle k, and continuous variables w_i^k, $i \in N$, $k \in K$, indicating the time at which vehicle k starts servicing vertex i. Let $\delta^+(i) = \{j\colon (i, j) \in A\}$ and $\delta^-(j) = \{i\colon (i, j) \in A\}$. The problem can then be stated as follows (see, e.g., Desrochers et al., 1988):

$$\text{minimize} \quad \sum_{k\in K} \sum_{(i,j)\in A} c_{ij} x_{ij}^k \tag{11}$$

subject to

$$\sum_{k\in K} \sum_{j\in\delta^+(i)} x_{ij}^k = 1, \quad i \in N, \tag{12}$$

$$\sum_{j\in\delta^+(0)} x_{0j}^k = 1, \quad k \in K, \tag{13}$$

$$\sum_{i\in\delta^-(j)} x_{ij}^k - \sum_{i\in\delta^+(j)} x_{ji}^k = 0, \quad k \in K, j \in N, \tag{14}$$

$$\sum_{i\in\delta^-(n+1)} x_{i,n+1}^k = 1, \quad k \in K, \tag{15}$$

$$x_{ij}^k\left(w_i^k + s_i + t_{ij} - w_j^k\right) \leqslant 0, \quad k \in K, (i, j) \in A, \tag{16}$$

$$a_i \leqslant w_i^k \leqslant b_i, \quad k \in K, i \in V, \tag{17}$$

$$\sum_{i\in N} q_i \sum_{j\in\delta^+(i)} x_{ij}^k \leqslant Q, \quad k \in K, \tag{18}$$

$$x_{ij}^k \in \{0,1\}, \quad k \in K, (i,j) \in A. \tag{19}$$

The objective function (11) minimizes the total routing cost. Constraints (12) state that each customer is visited exactly once, while constraints (13)–(15) ensure that each vehicle is used exactly once and that flow conservation is satisfied at each customer vertex. The consistency of the time variables w_i^k is ensured through constraints (16) while time windows are imposed by (17). These constraints also eliminate subtours. Finally, constraints (18) enforce the vehicle capacity restriction.

Formulation (11)–(19) is nonlinear because of constraints (16). These constraints can, however, be linearized as follows:

$$w_j^k \geqslant w_i^k + s_i + t_{ij} - M_{ij}\big(1 - x_{ij}^k\big), \quad k \in K, (i,j) \in A, \tag{20}$$

where $M_{ij} = \max\{0, b_i + s_i + t_{ij} - a_j\}$ is a constant. As suggested by Desrochers and Laporte (1991), the bounds on the time variables b_i^k can also be strengthened:

$$w_i^k \geqslant a_i + \sum_{j \in \delta^-(i)} \max\{0, a_j - a_i + s_j + t_{ji}\} x_{ji}^k, \quad k \in K, i \in V, \tag{21}$$

$$w_i^k \leqslant b_i - \sum_{j \in \delta^+(i)} \max\{0, b_i - b_j + s_i + t_{ij}\} x_{ij}^k, \quad k \in K, i \in V. \tag{22}$$

3.2 Exact algorithms for the VRPTW

As for most other vehicle routing problems, it is difficult to solve the VRPTW exactly through classical simplex-based branch-and-bound methods, even for small instances. This is in large part explained by the fact that the LP relaxation of the problem provides a weak lower bound. The first optimization algorithm for the VRPTW can be attributed to Kolen et al. (1987) who used dynamic programming coupled with state space relaxation (Christofides et al., 1981b) to compute lower bounds within a branch-and-bound algorithm. Instances with $n \leqslant 15$ were solved using this approach. Most subsequent algorithms rely either on the generation of valid inequalities to strengthen the LP relaxation or on mathematical decomposition techniques. This section reviews the three main available approaches: Lagrangian relaxation, column generation, and branch-and-cut. Additional references on the subject can also be found in the Cordeau et al. (2002a) review.

3.2.1 Lagrangian relaxation based algorithms

Lagrangian relaxation can be applied to the VRPTW in several ways. It is well known that when the subproblem obtained by relaxing some of the constraints possesses the integrality property, the best lower bound obtained by Lagrangian relaxation (i.e., the value of the Lagrangian dual) is equal to the value of the linear programming relaxation of the original problem. But

as mentioned above, the LP relaxation of formulation (11)–(19) provides a weak lower bound which will usually prevent the problem from being solved by branch-and-bound. As a result, successful implementations of Lagrangian relaxation for the VRPTW should retain at least some of the complicating constraints in the subproblem.

Fisher (1994) and Fisher et al. (1997) have described Lagrangian relaxation based on m-trees (see Section 2.2.1). This approach relaxes the flow conservation constraints as well as the capacity and time window constraints. Violated capacity constraints are handled by identifying subsets of customers $S \subseteq N$ that must be visited by at least $\kappa(S)$ vehicles and imposing the constraint

$$\sum_{k \in K} \sum_{i \in V \setminus S} \sum_{j \in S} x_{ij}^k \geqslant \kappa(S). \tag{23}$$

These constraints are relaxed in a Lagrangian fashion so that the resulting problem remains an m-tree problem with modified costs. Time windows are handled similarly by identifying infeasible paths and imposing the constraint that at least one arc in the path be left out of the solution. This approach has solved a few of the Solomon (1987) test instances with $n = 100$. In addition to the m-tree relaxation method, Fisher et al. (1997) have also experimented with a variable splitting approach in which additional variables y_i^k, equal to 1 if and only if customer i is visited by vehicle k, are introduced in the formulation, and the constraints $\sum_{j \in V} x_{ij}^k = y_i^k$ $(i \in N, k \in K)$ are dualized. The Lagrangian subproblem decomposes into a semi-assignment problem in the y_i^k variables which is solvable by inspection, and a set of m elementary shortest path problems with time windows and capacity constraints.

Another possible Lagrangian relaxation consists of dualizing the demand constraints. Let $\boldsymbol{\lambda} = (\lambda_i)$ $(i \in N)$ be the vector of multipliers associated with constraints (12) requiring that each customer be visited exactly once. For given values of the multipliers, the Lagrangian subproblem $L(\boldsymbol{\lambda})$ obtained by relaxing these constraints in the objective function is

$$\min \sum_{k \in K} \sum_{(i,j) \in A} (c_{ij} - \lambda_i)\, x_{ij}^k + \sum_{i \in N} \lambda_i, \tag{24}$$

subject to constraints (13)–(19).

This subproblem does not possess the integrality property. It does, however, decompose into m disjoint elementary shortest-path problems with capacity and time window constraints. When all vehicles are identical, a single problem can be solved to compute the lower bound. The Lagrangian dual, i.e., the problem of finding optimal multipliers that maximize $L(\boldsymbol{\lambda})$, is a concave nondifferentiable maximization problem. Using subgradient and bundle methods, Kohl and Madsen (1997) were able to solve some instances with up to 100 customers. They reported optimal solutions to each of the 27 clustered and short-horizon Solomon instances.

Kallehauge et al. (2006) have developed a stabilized cutting-plane algorithm to solve the Lagrangian dual. Cutting planes are generated by solving the Lagrangian subproblem and are introduced in a master problem which imposes bounds (i.e., a trust region) on the dual variables to ensure the stability of their values from one iteration to the next. Optimizing the relaxed master problem (a maximization linear program) provides a lower bound on the value of the original problem. To obtain feasible integer solutions, the cutting-plane algorithm is embedded within a branch-and-bound algorithm and valid inequalities are introduced in the master problem. Because the relaxed master problem is stated on the dual variables, violated subtour elimination constraints and 2-path inequalities (see Section 3.2.2) are added as columns to this problem. This approach has yielded good results on the Solomon test instances and was able to solve two large instances with 400 and 1000 customers, respectively.

3.2.2 Column generation algorithms

Column generation is intimately related to constraint generation and can be seen as a special way of updating the multipliers associated with the relaxed constraints. Let Ω^k denote the set of feasible paths for vehicle $k \in K$. For each path $\omega \in \Omega^k$, let c_ω^k be the cost of this path and let θ_ω^k be a binary variable equal to 1 if and only if vehicle k uses path ω. Let also $a_{i\omega}$ be the number of times customer $i \in N$ is visited by path ω. As first suggested by Balinski and Quandt (1964), the VRPTW can be stated as follows:

$$\text{minimize} \quad \sum_{k\in K} \sum_{\omega\in\Omega^k} c_\omega^k \theta_\omega^k \tag{25}$$

subject to

$$\sum_{k\in K} \sum_{\omega\in\Omega^k} a_{i\omega} \theta_\omega^k = 1, \quad i \in N, \tag{26}$$

$$\sum_{\omega\in\Omega^k} \theta_\omega^k = 1, \quad k \in K, \tag{27}$$

$$\theta_\omega^k \in \{0, 1\}, \quad k \in K, \omega \in \Omega^k. \tag{28}$$

Because the sets Ω^k are likely to have a very large cardinality, this problem can be tackled by a branch-and-bound algorithm in which the linear relaxations are solved by column generation. At each node of the enumeration tree, a restricted column generation master problem is solved over the current set of columns. New columns of negative reduced cost are generated by solving a resource constrained shortest path problem (13)–(19) with modified arc costs reflecting the current values of the dual variables associated with the constraints of the column generation master problem. This process stops when no negative reduced cost column can be generated. Because the column generation subproblem is equivalent to the Lagrangian subproblem $L(\boldsymbol{\lambda})$, the lower bound provided by column generation is equal to the value of the Lagrangian

dual. The dual of the LP relaxation of formulation (25)–(28) is, in fact, equivalent to the Lagrangian dual defined in the previous section. This formulation can also be obtained by applying the Dantzig–Wolfe decomposition principle (Dantzig and Wolfe, 1960) to the original formulation (11)–(19).

Branching must be performed at each node of the branch-and-bound tree, where the optimal solution to the linear relaxation includes fractional path variables. While it is in principle possible to branch directly on fractional θ_ω variables, this approach is difficult to implement in practice. Indeed, it is easy to set such variables equal to 1 but it is much more difficult to impose the opposite decision. In the latter case, care must be taken to ensure that the same path will not be generated more than once by the subproblem. To this purpose, one could use a modified dynamic programming algorithm to implicitly handle forbidden paths, or a p-shortest path algorithm where p is equal to the number of forbidden paths plus one. This would ensure the generation of at least one valid path of negative reduced cost whenever one exists. A more convenient branching scheme consists of making decisions on the original arc flow variables x_{ij}^k or on sums of these variables. For example, binary decisions can be made on the following sum of variables:

$$\sum_{j \in N'} \sum_{k \in K'} x_{ij}^k,$$

where $i \in N$, $N' \subseteq \delta^+(i)$, and $K' \subseteq K$. Forcing this sum to be equal to 1 requires that some vertex in subset N' be visited immediately after i by some vehicle. If $|N'| = 1$, then the corresponding vertex must be visited after i by some vehicle. If $|K'| = 1$, then vertex i is implicitly assigned to vehicle k. The special case $|N'| = 1$ and $|K'| = 1$ is equivalent to forcing $x_{ij}^k = 1$ for some given j and k. It is worth pointing out that all such decisions can be handled directly at the subproblem level through the simple elimination of arcs in the networks.

Column generation was successfully applied to the VRPTW by Desrochers et al. (1992) and by Kohl et al. (1999). The latter authors also used valid inequalities to strengthen the bounds obtained by column generation. More specifically, let

$$x(S) = \sum_{k \in K} \sum_{i \in V \setminus S} \sum_{j \in S} x_{ij}^k$$

denote the flow into set $S \subseteq N$ and denote by $\kappa(S)$ the minimum number of vehicles needed to serve all customers in S. Then the constraint

$$x(S) \geqslant \kappa(S) \tag{29}$$

is a valid inequality for the VRPTW and is called a κ-path inequality. Computing $\kappa(S)$ is a difficult problem which is equivalent to solving the VRPTW on a subset of vertices with the objective of minimizing the number of vehicles used. Kohl et al. (1999) have, in fact, restricted their attention to the case

$\kappa = 2$. Determining whether $\kappa(S) = 1$ for a particular subset S can be achieved by checking that the capacity of a single vehicle is sufficient and the corresponding TSPTW is feasible. The latter problem is NP-hard but can be solved relatively quickly by dynamic programming for small instances. The algorithm of Kohl et al. was capable of solving 70 of the 87 Solomon short-horizon instances to optimality. Cook and Rich (1999) have extended this approach to the case $\kappa \leqslant 6$ by using parallel computing and replacing the TSPTW feasibility problem with a VRPTW. They were thus able to solve 80 of the short-horizon instances. They also solved 30 of the 81 long-horizon instances.

While the constrained elementary shortest path problem is NP-hard, the relaxation obtained by allowing cycles can be solved by a pseudopolynomial labeling algorithm (see, e.g., Desrochers and Soumis, 1988). Because of time windows and capacity constraints, these cycles will nevertheless be of finite length. This relaxation will of course weaken the value of the lower bound, but cycle elimination procedures can be used to circumvent this difficulty. A procedure for eliminating 2-cycles (i.e., cycles of the form (i, j, i)) was first proposed by Houck et al. (1980). More recently, Irnich and Villeneuve (2003) developed an efficient approach to forbid cycles of length greater than 2. Experiments performed by the authors show that k-cycle elimination with $k \geqslant 3$ can substantially improve the lower bounds. Embedding this technique within column generation enabled the exact solution of 15 previously unsolved instances of the Solomon benchmark set.

Recently, Chabrier (2006) proposed a modified labeling algorithm to handle the constrained elementary shortest path problem and thus obtain improved lower bounds. In this algorithm, both exact and heuristic dominance rules are considered. Whenever the heuristic approach cannot find a path of negative reduced cost, the exact but slower implementation is used. This approach has allowed the author to find the optimal solution to 17 previously unsolved long-horizon instances from the Solomon benchmark set.

Promising results were also reported by Danna and Le Pape (2003) who developed a cooperation scheme between column generation and local search applied to the VRPTW. During the branch-and-price process, local search is regularly applied from the best known integer solution. This often results in an improved upper bound that can then be used to prune nodes in the enumeration tree. Furthermore, columns associated with solutions identified during local search can be fed into the restricted master problem. The branch-and-price algorithm thus benefits from local search by being provided at an early stage with high quality upper bounds, resulting in a smaller search tree. In turn, local search benefits from branch-and-price by working with a variety of different initial solutions, resulting in an effective form of diversification.

3.2.3 A branch-and-cut algorithm

A branch-and-cut algorithm for the VRPTW was developed by Bard et al. (2002). As in most such algorithms for the VRP, the problem is formulated using two-index variables x_{ij} equal to 1 if and only if a vehicle travels directly

from vertex i to vertex j. The algorithm incorporates five types of inequalities: subtour elimination constraints, capacity constraints, comb inequalities, incompatible pair inequalities, and incompatible path inequalities. At each node of the search tree an upper bound is computed by means of the Greedy Randomized Adaptive Search Procedure (GRASP) described by Kontoravdis and Bard (1995).

Incompatible pair inequalities rely on the existence of vertex pairs that cannot belong to the same vehicle route. If i and j denote two incompatible vertices and $\mathcal{P} = (i, h_1, \ldots, h_{|\mathcal{P}|-2}, j)$ is a path, then the following inequality is valid:

$$x_{i_1,h_1} + x_{h_1,i} + \cdots + x_{h_{|\mathcal{P}|-2},j} + x_{j,h_{|\mathcal{P}|-2}} \leqslant |\mathcal{P}| - 2. \tag{30}$$

Incompatible path inequalities are similar to infeasible pair inequalities but take arc orientations into account. If i and j are two vertices such that i cannot precede j in a feasible vehicle route then the following inequality is valid for any path $\mathcal{P}$ between i and j:

$$x_{i,h_1} + x_{h_1,h_2} + \cdots + x_{h_{|\mathcal{P}|-2},j} \leqslant |\mathcal{P}| - 2. \tag{31}$$

The authors present four separation heuristics to identify violated capacity constraints. The first is based on the computation of minimum cuts in G. The second applies a graph shrinking heuristic similar to that proposed by Araque et al. (1994) for the VRP. The third consists of identifying connected components in G that do not contain the depot. Finally, the fourth is a heuristic proposed by Kohl et al. (1999) to identify violated 2-path inequalities. Heuristic separation algorithms are also described for the identification of violated comb inequalities, incompatible path inequalities, and incompatible pair inequalities. The branch-and-cut algorithm of Bard, Kontoravdis, and Yu has obtained good results on the Solomon test instances: all 50-customer instances and several 100-customer instances were solved optimally.

3.3 Heuristics for the VRPTW

Because of the difficulty of the VRPTW and its high practical relevance, there is a genuine need to develop fast algorithms capable of producing good quality solutions in short computing times. Heuristics can also be used to provide upper bounds for the exact algorithms described in the previous section. This section describes the three main classes of heuristics for the VRPTW: construction heuristics, improvement heuristics, and metaheuristics.

3.3.1 Construction heuristics

Route construction algorithms work by inserting customers one at a time into partial routes until a feasible solution is obtained (see Section 2.3.1). Routes can either be constructed sequentially or in parallel. Construction algorithms are mainly distinguished by the order in which customers are selected and by the method used to determine where a customer should be inserted.

Several sequential insertion heuristics for the VRPTW were proposed by Solomon (1987). Among these heuristics, the most efficient, called I1, consists of first selecting the farthest customer from the depot as a seed customer. The remaining customers are then inserted one at a time into the current route by selecting at each iteration the customer that maximizes a saving measure, taking into account the distance from the depot and the cost of insertion in the current route. The customer is then inserted in the position minimizing a weighted combination of extra distance and extra time required to visit the customer. The process is repeated until all customers have been inserted or it is no longer possible to insert additional customers without violating either the capacity or time window constraints. At this point a new route is initialized by selecting a new seed customer and the process repeats itself until no customers remain.

A parallel version of this heuristic was later developed by Potvin and Rousseau (1993) who proposed a generalized regret measure to select the next customer for insertion. This measure reflects the cost increase likely to result if a customer is not assigned to the route minimizing the insertion cost. Further improvements to the sequential heuristic of Solomon (1987) were also described by Ioannou et al. (2001) who proposed modifying the criteria for customer selection and insertion to take into account the impact of the insertion on all routed and unrouted customers.

3.3.2 Improvement heuristics

Improvement heuristics iteratively improve an initial feasible solution by performing exchanges while maintaining feasibility. The process normally stops when no further exchange can be made without deteriorating the solution. Improvement heuristics are mainly characterized by the type of exchanges considered at each iteration. These define the neighborhood of a solution, i.e., the set of solutions reachable from the current solution by performing a single exchange.

The first improvement heuristics for the VRPTW (see, e.g., Russell, 1977; Baker and Schaffer, 1986) were adaptations of the 2-opt (Croes, 1958), 3-opt (Lin, 1965), and Or-opt (Or, 1976) edge exchange mechanisms originally introduced for the TSP. Because of time windows, checking whether a given exchange maintains feasibility of the solution can be rather time consuming. Starting with the work of Savelsbergh (1985), several attempts have been made to develop efficient implementations of neighborhood evaluation procedures for λ-exchanges (see also Solomon et al., 1988; Savelsbergh, 1990, 1992). A comparison of 2-opt, 3-opt, and Or-opt exchange heuristics for the VRPTW was performed by Potvin and Rousseau (1995) who also introduced a new exchange, called 2-opt*, a special case of 2-opt that maintains the orientation of the subroutes involved in the exchange. This is accomplished by removing the last n_1 customers from a route k_1, inserting them after the first n_2 customers of a route k_2, and reconnecting the initial part of route k_1 with the terminal part of route k_2. Another exchange mechanism was described by Thompson and

Psaraftis (1993) who proposed transferring sets of customers in a cyclic fashion between routes.

Several attempts have also been made to integrate construction and improvement heuristics. Russell (1995) developed a procedure that embeds route improvement within the solution construction process. More precisely, customers can be switched between routes, and routes can be eliminated during the construction of the solution which is performed by a procedure similar to that of Potvin and Rousseau (1993). More recently, Cordone and Wolfler Calvo (2001) have proposed a composite heuristic in which a set of initial solutions is first constructed by means of Solomon's I1 insertion heuristic and an improvement procedure is then applied to each of them. This procedure applies 2-opt and 3-opt exchanges and attempts to reduce the number of routes by relocating customers. To escape from local optima, the heuristic alternates between an objective minimizing total distance and an objective minimizing total route duration (the primary objective being in both cases to minimize the number of routes). Several deterministic local search heuristics were also proposed by Bräysy (2002), based on a new three-phase approach. In a first phase, an initial solution is created with one of two proposed route construction heuristics (a cheapest insertion-based heuristic with periodic route improvements and a parallel savings heuristic). The second phase attempts to reduce the number of routes by applying a local search operator based on ejection chains (see, e.g., Glover, 1992). Finally, the third phase applies Or-opt exchanges to reduce the total length of the routes.

3.3.3 Metaheuristics

Most of the recent research on approximate algorithms for the VRPTW has concentrated on the development of metaheuristics. Unlike classical improvement methods, metaheuristics usually incorporate mechanisms to continue the exploration of the search space after a local minimum is encountered.

Tabu search heuristics. Some of the first applications of tabu search to the VRPTW can be attributed to Semet and Taillard (1993) and to Potvin et al. (1996) who combined Solomon's insertion heuristics with improvement schemes based on vertex and chain exchange procedures.

A more sophisticated algorithm was later developed by Taillard et al. (1997) for the VRP with soft time windows in which vehicles are allowed to arrive late at customer locations but time window violations are penalized in the objective function. This heuristic relies on the concept of adaptive memory introduced by Rochat and Taillard (1995) and on the decomposition and reconstruction procedure developed by Taillard (1993) for the classical VRP. An adaptive memory is a pool of routes extracted from the best solutions found during the search. This memory is first initialized with routes produced by a randomized insertion heuristic. At each iteration of the metaheuristic, a solution is constructed from the routes belonging to the adaptive memory and is improved through tabu search. The routes of the resulting solution are then stored in

the adaptive memory if this solution improves upon the worst solution already stored. The tabu search heuristic uses an exchange operator, called CROSS exchange, which swaps sequences of consecutive customers between two routes. Individual routes are also optimized by removing two edges from a route and moving the segment between these two edges to another location within the route. A parallel computing implementation of this approach is described in Badeau et al. (1997).

A metaheuristic embedding reactive tabu search (see, e.g., Battiti and Tecchiolli, 1994) within the parallel construction heuristic of Russell (1995) was developed by Chiang and Russell (1997). In this implementation, the tabu list length is increased if identical solutions occur too frequently and is decreased if no feasible solution can be found. Using a variety of customer ordering rules and criteria for measuring the best insertion points, the metaheuristic first constructs six different initial solutions by gradually inserting customers and repeatedly applying tabu search to the partial solutions. The best solution obtained after this step is further improved through tabu search. Exchanges are performed by using some of the λ-interchanges of Osman (1993): switch a customer from one route to another and swap two customers belonging to different routes.

More recently, a tabu search heuristic was developed by Cordeau et al. (2001) for the VRPTW and two of its generalizations: the periodic VRPTW and the multidepot VRPTW (see also Cordeau et al., 1997). In this heuristic, an initial solution is obtained by means of a modified sweep heuristic. Infeasible solutions are allowed during the search and violations of capacity, duration or time window constraints are penalized in the objective function through dynamically updated penalty factors. At each iteration of the tabu search, a customer is removed from its current route and inserted into a different route by using a least cost insertion criterion. A continuous diversification mechanism that penalizes frequently made exchanges is used to drive the search process away from local optima. Finally, a post-optimizer based on a specialized TSPTW heuristic (Gendreau et al., 1998) is applied to individual routes. An improvement to this heuristic for the handling of route duration constraints was recently described by Cordeau et al. (2004). The heuristic was also extended by Cordeau and Laporte (2001) to handle heterogeneous vehicles. Other tabu search algorithms for the VRPTW were proposed by Brandão (1998), Schulze and Fahle (1999), and Lau et al. (2003).

Genetic algorithms. Homberger and Gehring (1999) have described two evolution strategies for the VRPTW. Both are based on the (μ, λ) strategy: starting from a population with μ individuals, subsets of individuals are randomly selected and recombined to yield a total of $\lambda > \mu$ offspring. Each offspring is then subjected to a mutation operator, and the μ fittest are selected to form the new population. In the first method, new individuals are generated directly through mutations and no recombination takes place. Mutations are obtained by performing one or several moves from the 2-opt, Or-opt, and

1-interchange families. In the second method, offspring are generated through a two-step recombination procedure in which three individuals are involved. In both methods, the fitness of an individual depends first on the number of vehicles used, and second on the total distance traveled. Gehring and Homberger (2002) later proposed a two-phase metaheuristic in which the first phase minimizes the number of vehicles through an evolution strategy, while the second one minimizes the total distance through tabu search. A parallelization strategy is also used to run several concurrent searches of the solution space with differently configured metaheuristics cooperating through the exchange of solutions.

Berger et al. (2003) have developed a genetic algorithm that concurrently evolves two distinct populations pursuing different objectives under partial constraint relaxation. The first population aims to minimize the total distance traveled while the second one focuses on minimizing the violations of the time window constraints. The maximum number of vehicles imposed in the first population is equal to $k_{\min}$ whereas the second population is allowed only $k_{\min} - 1$ vehicles, where $k_{\min}$ refers to the number of routes in the best known feasible solution. Whenever a new feasible solution emerges from the second population, the first population is replaced with the second and the value of $k_{\min}$ is updated accordingly. Two recombination operators and five mutation operators are used to evolve the populations. This approach has proved to be rather efficient in minimizing the number of vehicles used.

More recently, Mester and Bräysy (2005) have developed an iterative metaheuristic that combines guided local search and evolution strategies. An initial solution is first created by an insertion heuristic. This solution is then improved by the application of a two-stage procedure. The first stage consists of a guided local search procedure in which 2-opt* and Or-opt exchanges are performed together with 1-interchanges. This local search is guided by penalizing long arcs appearing often in local minima. The second stage iteratively removes a selected set of customers from the current solution and reinserts the removed customers at minimum cost. These two stages are themselves repeated iteratively until no further improvement can be obtained. Very good results are reported by the authors on large-scale instances. According to Bräysy and Gendreau (2005b), the three approaches just described seem to produce the best results among genetic algorithms. Other such algorithms have also been proposed by a number of researchers including Potvin and Bengio (1996), Thangiah and Petrovic (1998), and Tan et al. (2001).

Other metaheuristics. Kontoravdis and Bard (1995) have described a two-phase GRASP for the VRPTW. A number of routes are first initialized by selecting seed customers. The remaining customers are then gradually inserted in the routes by using a randomized least insertion cost procedure. During this process, periodic attempts are made to improve the routes by local search. In this phase certain routes may be eliminated by means of a deterministic procedure that attempts to relocate the customers to a different route. To estimate

the required number of routes, the authors have proposed three lower bounds for fleet size. Two are based on bin packing structures generated by the capacity or time window constraints. The other is derived from the associated graph created by pairs of customers having incompatible demands or time windows.

A guided local search algorithm for the VRPTW was introduced by Kilby et al. (1998). In guided local search, the objective function is augmented with a penalty term reflecting the proximity of the current solution value to that of previously encountered local minima. The method is used to drive a local search heuristic that modifies the current solution by performing one of four moves: 2-opt exchanges within a route, switching a customer from one route to another, exchanging customers belonging to two different routes, and swapping the ends of two routes. All customers are first assigned to a virtual vehicle and the routes for the actual vehicles are left empty. Because a penalty is associated with not visiting a customer, a feasible solution will be constructed in the process of minimizing cost. The local search algorithm starts from this solution and performs a series of exchanges until a local minimum is reached. The objective function is then modified by adding a term penalizing the presence of the arcs used in this solution. The search iterates by finding new local minima and accumulating penalties until a stopping criterion is met. This approach was later coupled with tabu search and embedded within a constraint programming framework by De Backer et al. (2000).

Gambardella et al. (1999) have developed an ant colony optimization algorithm for the VRPTW which associates an attractiveness measure to the arcs. Artificial ants represent parallel processes whole role is to construct feasible solutions. To deal with the hierarchical objective of first minimizing the number of vehicles and then minimizing distance, two ant colonies are used, each dedicated to the optimization of a different objective. These colonies cooperate by exchanging information through pheromone updating. Whenever a feasible solution with a smaller number of vehicles is found, both colonies are reactivated with the reduced number of vehicles.

Bent and Van Hentenryck (2004) have described a two-stage hybrid algorithm that first minimizes the number of routes by simulated annealing and then minimizes total distance traveled by using a large neighborhood search (Shaw, 1998) which may relocate a large number of customers. The first stage uses a lexicographic evaluation function to minimize the number of routes, maximize the sum of the squares of the route sizes, and minimize the minimal delay (a measure of time window tightness) of the solution. The neighborhood used in this stage consists of 2-opt, Or-opt, relocating, exchange, and crossover moves. In the second stage, subsets of customers are removed from their current route and reinserted in possibly different routes. Customers selected for removal are randomly chosen but the algorithm favors customers that are geographically close to each other and belong to different routes. A branch-and-bound algorithm is then used to reinsert these customers.

A four-phase metaheuristic based on a modification of the variable neighborhood search was described by Bräysy (2003). In the first phase, an initial

solution is created by using route construction heuristics. During this process, the partial routes are periodically reoptimized through Or-opt exchanges. In the second phase, an attempt is made to reduce the number of routes by applying a route elimination operator based on ejection chains. In the third phase, four local search procedures embedded within a variable neighborhood search (see, e.g., Mladenović and Hansen, 1997) are applied to reduce the total distance traveled. These procedures are based on modifications to the CROSS exchanges of Taillard et al. (1997) and cheapest insertion heuristics. In the fourth phase, a modified objective function considering waiting time is used by the local search operators in the hope of further improving the solution.

More recently, a local search algorithm with restarts was also proposed by Li and Lim (2003). This algorithm first constructs an initial solution by using an insertion heuristic. Local search is then performed from this solution using three exchange operators that move segments of customers either between routes or within the same route. Whenever a local minimum is reached, multiple restarts are performed starting from the best known solution, and a tabu list is used to prevent cycling.

A large number of other metaheuristics based on various paradigms have been described in recent years. For additional references on approximate algorithms for the VRPTW as well as detailed computational experiments, the reader is referred to recent surveys by Bräysy and Gendreau (2005a, 2005b).

4 The inventory routing problem

The Inventory Routing Problem (IRP) is an important extension of the VRP which integrates routing decisions with inventory control. The problem arises in environments where Vendor Managed Inventory (VMI) resupply policies are employed. These policies allow a vendor to choose the timing and size of deliveries. In exchange for this freedom, the vendor agrees to ensure that its customers do not run out of product. In a more traditional relationship, where customers call in their orders, large inefficiencies can occur due to the timing of customers' orders (resulting in high inventory and distribution costs). Realizing the cost savings opportunities of vendor managed inventory policies, however, is not a simple task, particularly with a large number and variety of customers. The inventory routing problem achieves this goal by determining a distribution strategy that minimizes *long term* distribution costs. This description of the inventory routing problem focuses primarily on distribution. Inventory control is restricted to ensuring that no stockouts occur at the customers. Inventory control takes a more prominent role when inventory holding costs are considered. In the inventory control literature, the resulting environment is usually referred to as a *one warehouse multiretailer system*.

Inventory routing problems are very different from VRPs. Vehicle routing problems occur when customers place orders and the vendor, on any given day,

assigns the orders for that day to routes for vehicles. In IRPs, the delivery company, not the customer, decides how much to deliver to which customers each day. There are no customer orders. Instead, the delivery company operates under the restriction that its customers are not allowed to run out of product. Another key difference is the planning horizon. Vehicle routing problems typically deal with a single day, the only requirement being that all orders have to be delivered by the end of the day. Inventory routing problems are defined on a longer horizon. Each day the delivery company makes decisions about which customers to visit and how much to deliver to each of them, while keeping in mind that decisions made today have an impact on what has to be done in the future. The objective is to minimize the total cost over the planning horizon while ensuring that no customer runs out of product.

4.1 Definition of the IRP

The deterministic IRP is concerned with the repeated distribution of a single product from a single facility, to a set of n customers over a planning horizon of length T, possibly infinity. Customer i consumes the product at a rate u_i (say volume per day) and can maintain a local inventory of product of up to a maximum of C_i. The inventory at customer i is I_i^0 at time 0. A fleet of m homogeneous vehicles, with capacity D, is available for the distribution of the product. If a quantity d_i is delivered at customer i, the vendor earns a reward equal to $r_i d_i$. It takes a vehicle a time t_{ij} to traverse arc (i, j) of the distribution network and a cost c_{ij} is incurred when doing so. The objective is to maximize the profit (revenue minus cost) over the planning horizon, without causing stockouts at any of the customers. (Note that because product usage is assumed to be deterministic and no stockouts are allowed, long run revenues are fixed and the key is to reduce delivery costs.) A dispatcher has to decide when to serve a customer, how much to deliver, and which delivery routes to use to serve customers.

In the stochastic IRP customer demands are defined at discrete time instants t by means of random variables. Let $U_t = (U_{1t}, \ldots, U_{nt})$ denote the vector of random customer demands at time t. Customer demands on different days are independent random vectors with a joint probability distribution F that does not change with time; that is, $U_0, U_1, \ldots$ is an independent and identically distributed sequence, and F is the probability distribution of each U_t. The probability distribution F is known to the decision maker. The vendor can measure the inventory level X_{it} of each customer i at any time t. At each time instant t, the vendor makes a decision that controls the routing of vehicles and the replenishment of customer inventories. Because demand is uncertain, there is often a positive probability that a customer will run out of stock, and thus shortages cannot always be prevented. Shortages result in a penalty $p_i s_i$ if the unsatisfied demand on day t at customer i is s_i. Unsatisfied demand is treated as lost demand. The objective is to construct a distribution policy maximizing the expected discounted profit over an infinite time horizon.

4.2 Motivating example

To illustrate the difficulty of inventory routing problems, we reproduce a small deterministic example introduced by Fisher et al. (1982) and Bell et al. (1983). The relevant optimal tour costs can be derived from the network shown in Figure 1, e.g., the optimal tour costs for visiting customers 1 and 2, denoted by $C_{1,2}$, is equal to \$210. The vehicle capacity is 5000 gallons and customer tank capacity and usage data, in gallons, are as follows:

Customer i	d_i	u_i
1	5000	1000
2	3000	3000
3	2000	2000
4	4000	1500

A simple schedule jointly replenishes customers 1 and 2 as well as customers 3 and 4 on a daily basis. This schedule is natural because customers 1 and 2 (3 and 4, respectively) are near each other. Each customer i receives a quantity equal to its daily consumption u_i. The long-run average cost of this schedule is 420 miles per day. An improved schedule consists of a cycle that repeats itself every two days. On the first day, one trip replenishes 3000 gallons to customer 2 and 2000 gallons to customer 3, at a cost of 340 miles. On the second day, two trips are made. The first trip replenishes 2000 gallons to customer 1 and 3000 gallons to customer 2. The second trip replenishes 2000 gallons to

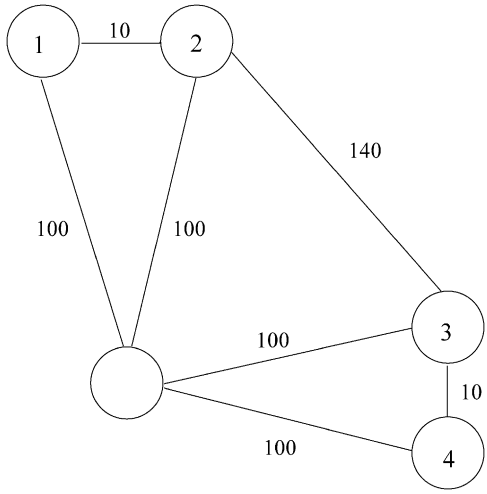


Fig. 1. A four-customer example with distances shown on edges.

customer 3 and 3000 gallons to customer 4. Each trip costs 210 miles. The average cost of this schedule is 380 miles per day, which is nearly 10% lower than the first schedule.

4.3 Observations on the IRP

Before describing solution approaches, we present some general observations concerning inventory routing problems and some common elements found in most solution approaches.

The IRP is a long-term dynamic control problem which is extremely difficult to solve. Therefore, most of the available algorithms solve only a short-term planning problem. In early publications, it was often just a single day but later, short-term was expanded to a few days. Two key issues need to be resolved with such approaches: how to model the long-term effect of short-term decisions, and which customers to consider in the short-term planning period. A short-term approach that only minimizes costs has the tendency to defer as many deliveries as possible to future planning periods, which may lead to an undesirable situation in the future. Therefore, a proper incorporation of the long-term objective into the short-term planning problem is essential. The long-term effect of short-term decisions needs to capture the costs and benefits of delivering to a customer earlier than necessary. This usually means delivering less and may lead to higher future distribution costs, but reduces the risk of a stockout and may thus reduce future shortage costs. Decisions regarding which customers need to be considered in the short-term planning period are usually guided by some measure of the urgency to make a delivery to a customer and the quantity that can be delivered. Usually, it is assumed that customers considered in the short-term planning period may actually be visited, but the decision whether or not to actually visit them still has to be made.

When the short-term planning problem consists of a single day, the problem can be viewed as an extension of the VRP and solution techniques for the VRP can be adapted. For example, Campbell and Savelsbergh (2004c) have discussed efficient implementations of insertion heuristics to handle situations where the delivery amount has to lie between a lower and an upper bound, as opposed to being fixed. In related work, Campbell and Savelsbergh (2004b) have studied the problem of determining an optimal delivery schedule for a route, i.e., given a sequence of customer visits, determine the timing of the visits so as to maximize the total amount of the product delivered on the route. Because single day approaches usually base decisions on the latest inventory measurement and a predicted usage for that day, they avoid the difficulty of forecasting long-term usage, which makes the problem much simpler.

4.4 Single customer analysis

It is insightful to analyze the "simple" situation in which there is only a single customer. The results of this type of analysis can be used effectively to guide

decisions on which customers to consider in a short term planning horizon. The material presented in this subsection is primarily based on Jaillet et al. (2002), although much of it dates back to the work of Dror and Ball (1987). We first consider the deterministic case. For ease of notation, let the usage rate of the customer be u, the storage capacity of the customer be C, the initial inventory level be I^0, the delivery cost to the customer be c, and the vehicle capacity be Q. It is easy to see that an optimal policy is to fill up the storage space precisely at the time when it becomes empty. Therefore the cost v_T for a planning period of length T is

$$v_T = \max\left\{0, \left\lceil \frac{Tu - I^0}{\min\{C, Q\}} \right\rceil\right\} c.$$

Now consider the stochastic case in which one decides daily whether to make a delivery to the customer or not. The demand U between consecutive decision points, i.e., the demand per day, is a random variable with known probability distribution and finite mean. Assuming that the storage capacity at each customer is at least as large as the vehicle capacity and the vendor can only monitor the inventory in the storage space at the time of a delivery, it can be shown that for the infinite horizon case, there exists an optimal policy that fills up the storage space at each delivery and, following any scheduled or stockout delivery, plans the next delivery d days after. The optimal replenishment interval d is a constant chosen to minimize the expected daily cost.

A d-day policy makes a delivery to the customer every d days and delivers as much as possible, unless a stockout occurs earlier. In such a case, the vehicle is sent right away, which generates a cost S. It is assumed that deliveries are instantaneous, so that no additional stockout penalties are incurred. Furthermore, assume that initially the storage space is full. Let p_j be the probability that a stockout first occurs on day j ($1 \leqslant j \leqslant d-1$). Then $p = p_1 + p_2 + \cdots + p_{d-1}$ is the probability that there is a stockout in period $[1, \ldots, d-1]$. Furthermore, let $v_T(d)$ be the expected total cost of this policy over a planning period of length T. We now have for $d > T$

$$v_T(d) = \sum_{1 \leqslant j \leqslant T} p_j \big(v_{T-j}(d) + S\big)$$

and for $d \leqslant T$

$$v_T(d) = \sum_{1 \leqslant j \leqslant d-1} p_j \big(v_{T-j}(d) + S\big) + (1-p)\big(v_{T-d}(d) + c\big).$$

As a consequence, the expected total cost of filling up a customer's tank every d days over a T-day period ($T \geqslant d$) is given by

$$v_T(d) = \alpha(d) + \beta(d)T + f(T, d),$$

where $\alpha(d)$ is a constant depending only on d, $f(T, d)$ is a function that tends to zero exponentially fast as T tends to infinity, and

$$\beta(d) = \frac{pS + (1-p)c}{\sum_{1 \leqslant j \leqslant d} jp_j},$$

with $p_d = 1 - p$. The value $\beta(d)$ is the long-run average cost per day. To determine the best policy in this class, we need to minimize $v_T(d)$ which for large T means finding a value of d minimizing $\beta(d)$.

4.5 *The two-customer IRP*

When more than one customer is served, the problem becomes significantly harder. Not only is it necessary to decide which customers to visit next, but one must also determine how to combine them into vehicle tours, and how much to deliver to each of them. Even if there are only two customers, these decisions may not be easy. The material in the remainder of this section is primarily based on Campbell et al. (1998).

If the two customers are visited together, it is intuitively clear that given the amount delivered at the first customer, it is optimal to deliver as much as possible at the second one (determined by the remaining amount in the vehicle, and the remaining capacity at the second customer). Thus the problem of deciding how much to deliver to each customer involves a single decision. However, making that decision may not be easy, as the following two-customer stochastic IRP example shows.

Assume the product is delivered and consumed in discrete units and that each customer has a storage capacity of 20 units. The daily demands of the customers are independent and identically distributed (across customers as well as across time), with $P(U = 0) = 0.4$ and $P(U = 10) = 0.6$. The shortage penalty is $s_1 = 1000$ per unit at customer 1 and $s_2 = 1005$ per unit at customer 2. The vehicle capacity is 10 units. At the beginning of each day the inventory at the two customers is measured, and the decision maker determines how much to deliver to each customer. There are three possible vehicle tours, namely tours exclusively to customers 1 and 2, of cost 120 each, and a tour to both customers 1 and 2, of cost 180. Only one vehicle tour can be completed per day. This situation can be modeled as an infinite horizon Markov decision process, with the objective of minimizing the expected total discounted cost. Because of the small size of the state space, it is possible to compute the optimal expected value and an optimal policy.

Figure 2 shows the expected value (total discounted cost) as a function of the amount delivered at customer 1 (and therefore also at customer 2), when the inventory at each customer is 7, and both customers are to be visited in the next vehicle tour (which is the optimal decision in the given state). The figure shows that the objective function is not unimodal, with a local minimum at 3, and a global minimum at 7. Consequently, deciding just how much to deliver to each customer may require solving a nonlinear optimization problem with

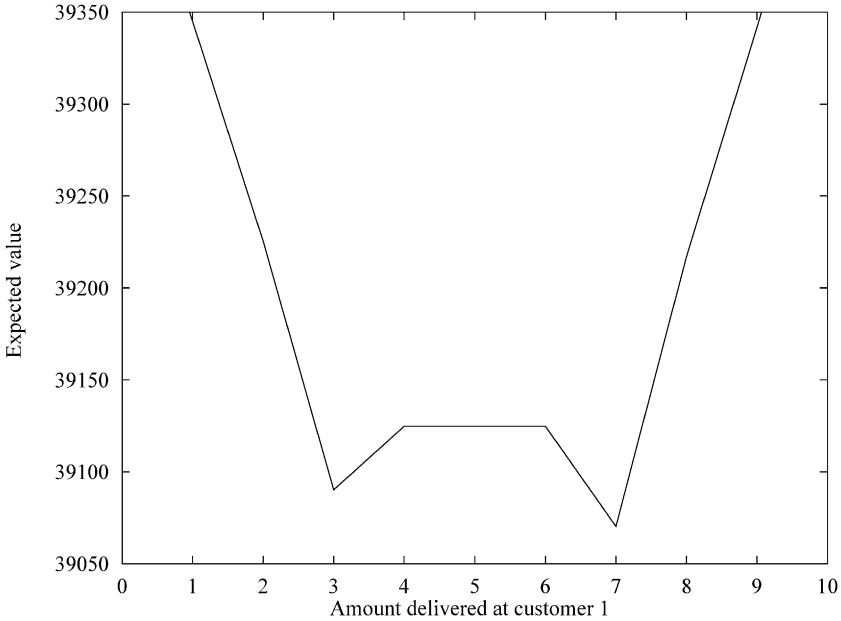


Fig. 2. Nonunimodal objective function for determining the optimal delivery quantity.

a nonunimodal objective function. This is a hard problem for which available search methods may not converge to an optimal solution.

4.6 Literature review on the IRP

Rather than providing a comprehensive review of the IRP literature, we discuss several research streams representing a variety of solution approaches that have been proposed and investigated. We encourage the reader to examine the referenced papers for more elaborate and precise coverage.

A first stream of research uses time-discretized integer programming models to determine the set of customers to be visited in a short-term planning horizon as well as the amount of product to deliver to them. In order to accurately reflect costs and time related aspects, the integer linear programs work with a set of potential delivery routes. Fisher et al. (1982) and Bell et al. (1983) pioneered this approach when they studied the IRP at Air Products, a producer of industrial gases. Their formulation determines the delivery volumes to customers, the assignment of customers to routes, the assignment of vehicle to routes, and the start times of routes. The core structure of their model is presented below, where the variable x_{irtv} represents the amount of product delivered to customer $i \in N$ on route $r \in R$ starting at time $t \in T$, the variable y_{rt} is 1 if route r starts at time t and 0 otherwise, S_r the set of customers visited on route r, p_i the value of delivering a unit of product to customer i, F_r the fixed cost of executing route r, $\underline{q}_{it}$ a lower bound on the cumulative amount

delivered to customer i by time t, and $\bar{q}_{it}$ an upper bound on the cumulative amount that can be delivered to customer i by time t:

$$\text{maximize} \quad \sum_{r\in R}\sum_{t\in T}\left(\sum_{i\in S_r} p_i x_{irt} - F_r y_{rt}\right)$$

subject to

$$\underline{q}_{it} \leqslant \sum_{r\in R}\sum_{s\leqslant t} x_{irs} \leqslant \bar{q}_{it}, \quad i \in N, t \in T,$$

$$\sum_{i\in S_r} x_{irt} \leqslant Q y_{rt}, \quad r \in R, t \in T,$$

$$x_{irt} \geqslant 0, \quad y_{rt} \in \{0, 1\}.$$

In the model, the per unit value of a delivery to a customer is used to represent the effect of decisions on events occurring beyond the planning horizon of the model. In the short-term planning period considered by the model, there is considerable discretion in the amount of product to deliver. In the long run this amount is determined by customer usage. Hence, each unit scheduled for delivery to a customer within the planning horizon reduces the amount to be delivered in the future. This is accounted for by setting the unit value to an estimate of the cost of delivering to a customer at a point in time outside the planning horizon of the model. Furthermore, rather than explicitly incorporating customer usage rates into the model, lower and upper bounds on the cumulative amount to be delivered to each customer in each time period in the planning horizon are used. It is simple, of course, to convert customer usage rates into bounds, i.e., $\underline{q}_{it} = \max\{0, tu_i - I_i^0\}$ and $\bar{q}_{it} = tu_i + C_i - I_i^0$. Lagrangian relaxation was a central tool in developing an effective heuristic for solving the integer program. The size of the integer programs to be solved depends on the chosen time discretization as well as on the size of set of routes.

Campbell and Savelsbergh (2004a) use an integer linear program with a similar structure to determine which customers to visit in the next few days (even though the integer program covers several weeks) and to *suggest* quantities and delivery times to these customers. However, then, in a second phase, they use modified insertion heuristics to determine the actual delivery routes and quantities. The advantage of such a two-phase approach is that a higher degree of accuracy (in terms of timing of events) can be provided in the second phase and other practical details, such as drivers shifts, can be considered. The delivery quantities and times specified by the solution to the integer program are good from a long-term perspective; they may need to be modified somewhat to also be good from a short-term perspective. When constructing the actual delivery routes, Campbell and Savelsbergh consider delivering more to the customers than the quantity suggested by the integer program (and slightly altering the delivery time if needed) since this may result in higher vehicle utilization and thus higher revenues. As in Bard et al. (1998b) their approach is embedded into a rolling horizon framework.

A second stream of research is based on the single customer analysis presented above. This approach was pioneered by Dror et al. (1985) and Dror and Ball (1987). The optimal replenishment day t_i^* minimizing the expected total cost for customer i is used to determine the set of customers considered in a short-term planning problem for the next $\bar{t}$ days. If $t_i^* \leqslant \bar{t}$, then the customer will be included and will definitely be visited. A value c_t is computed for each day of the planning period to reflect the expected increase in future cost if the delivery is made on day t instead of t^*. If $t^* \geqslant T$, i.e., the optimal replenishment day falls outside the short-term planning period, then a future benefit g_t can be computed for making an early delivery to the customer on day t of the short-term planning period. These computed values reflect the long term effects of short term decisions. An integer linear program is subsequently solved to assign customers to a vehicle and a day, or just a day, that minimizes the sum of these costs plus the transportation costs. (It was shown by Adelman (2004) that this objective function is in fact equivalent to that used by Fisher et al. (1982).) The delivery amount to a customer on a specific day is fixed and set to the quantity needed to fill up the storage tank on that day. This leaves either TSPs or VRPs to be solved in the second stage. These ideas are extended and improved in Trudeau and Dror (1992). The most recent work along these lines is that of Bard et al. (1998a, 1998b) who work with a rolling horizon approach in which a short term planning problem is defined for a two-week period, but only the decisions for the first week are implemented. In addition, satellite facilities are considered, i.e., locations other than the depot where vehicles can be refilled.

A third stream of research focuses on the asymptotic analysis of delivery policies. Anily and Federgruen (1990, 1991, 1993) analyze fixed partition policies for the IRP with an unlimited number of vehicles. Customers within the same partition are divided into regions so as to make the demand of each region roughly equal to a vehicle load. A customer may appear in more than one region, but then a certain percent of his demand is allocated to each region. When one customer in a region is also visited, all other customers in that region are also visited. The authors determine lower and upper bounds on the minimum long-run average cost over all fixed partition policies, and propose a heuristic, called modified circular regional partitioning, to choose a fixed partition. Using similar ideas, Gallego and Simchi-Levi (1990) evaluate the long-run effectiveness of direct deliveries (one customer on each route). Direct shipping is shown to be at least 94% effective over all inventory routing strategies whenever the minimal economic lot size is at least 71% of vehicle capacity. This indicates that direct shipping becomes an undesirable and costly policy when many customers require significantly less than a vehicle load, making more complicated routing policies the appropriate choice. Another adaptation of these ideas can be found in Bramel and Simchi-Levi (1995) who consider a variant of the IRP in which customers can hold an unlimited amount of inventory. To obtain a solution, they transform the problem into a Capacitated Concentrator Location Problem (CCLP), solve it, transform the solution back

into a solution to the IRP, and heuristically improve it. The CCLP solution will partition the customers into disjoint sets, which in the IRP will become the fixed partitions. Chan et al. (1998) analyze zero-inventory ordering policies, in which a customer's inventory is replenished only when it has been depleted, in combination with fixed partitioning routing policies and derive asymptotic worst-case bounds on their performance. Gaur and Fisher (2004) consider an IRP with time varying demand. They propose a randomized heuristic to find a fixed partition policy with periodic deliveries. Their method was implemented for a supermarket chain.

The fourth stream of research is based on formulating the stochastic IRP as a Markov decision process and thus explicitly incorporating demand uncertainty. This approach was pioneered by Minkoff (1993) who proposed a decomposition heuristic to overcome the computational difficulties caused by large state spaces. The heuristic solves a linear program to allocate joint transportation costs to individual customers and then solves individual customer subproblems. The value functions of the subproblems are added to approximate the value function of the original problem. The main limitation of the proposed approach is that it assumes the availability of a set of delivery routes with fixed delivery quantities for the customers on a route and the dispatcher only has to decide which of the delivery routes to use at each decision point. This limitation is removed in the work of Kleywegt et al. (2002, 2004) on approximate dynamic programming approaches and in that of Adelman (2003a, 2004) on price-directed approaches. Let state $x = (x_1, x_2, \ldots, x_n)$ represent the current inventory at each customer, and let $\mathcal{A}(x)$ denote the set of all feasible decisions when the process is in state x. A decision $a \in \mathcal{A}(x)$ specifies which customer inventories to replenish, how much to deliver at each customer location, and how to combine customers into vehicle routes. Let Q be the Markov transition function according to which transitions occur. Let $g(x, a)$ denote the expected single stage net reward if the process is in state x at time t and decision $a \in \mathcal{A}(x)$ is implemented. The objective is to maximize the expected total discounted value over an infinite horizon. Let $V^*(x)$ denote the optimal expected value given that the initial state is x. Then, for any state x,

$$V^*(x) = \sup_{a \in \mathcal{A}(x)} \left\{ g(x, a) + \alpha \int V^*(y) Q[\mathrm{d}y|x, a] \right\}. \tag{32}$$

A policy π^* is called optimal if $V^{\pi^*} = V^*$, where V^π represents the value function of policy π. Solving a Markov decision process involves computing the optimal value function V^* and an optimal policy π^* by solving the optimality equation (32). This requires performing the following major computational tasks:

(1) The computation of the optimal value function V^*. Most algorithms for computing V^* involve the computation of successive approximations to $V^*(x)$ for every state x. These algorithms are practical only if the state space is small.

(2) The estimation of the expected value (the integral in (32)). For the stochastic IRP, this is a high dimensional integral. Conventional numerical integration methods are not practical for the computation of such high-dimensional integrals.
(3) The maximization problem on the right-hand side of (32) has to be solved to determine the optimal decision for each state. For the stochastic IRP, this means solving a complex variant of the VRP.

Kleywegt, Nori, and Savelsbergh develop approximation methods to efficiently perform these computational tasks. Furthermore, their approach has the ability to handle a finite fleet of vehicles, whereas in other Markov decision process based approaches it is assumed that there exists an infinite fleet of vehicles. The optimal value function V^* is approximated by $\widehat{V}$ as follows. First, the stochastic IRP is decomposed into subproblems defined for specific subsets of customers. Each subproblem is also a Markov decision process. The subsets of customers do not necessarily partition the set of customers, but must cover it. The idea is to define each subproblem so that it provides an accurate representation of the overall process as experienced by the subset of customers. To do so, the parameters of each subproblem are determined by simulating the overall stochastic IRP process, and by constructing simulation estimates of subproblem parameters. Next, each subproblem is solved optimally. Finally, for any given state x, the approximate value $\widehat{V}(x)$ is determined by choosing a partition of the customers and by setting $\widehat{V}(x)$ equal to the sum of the optimal value functions of the subproblems corresponding to the partition at states corresponding to x. The partition is chosen to maximize $\widehat{V}(x)$. Randomized methods, incorporating variance reduction techniques to limit the required sample size, are used to estimate the expected value on the right-hand side of (32). Action determination involves deciding which customers to visit on a route and how much to deliver to them. This is achieved through a heuristic. An initial solution consisting of only direct delivery routes is constructed. This is followed by a local search procedure that examines the benefit of adding a customer to an existing route and modifying the delivery quantities. Using their approach Kleywegt, Nori, and Savelsbergh can solve problems involving up to 50 customers.

More recently, Adelman (2003a, 2004) proposed a price-directed operating policy based on a simple economic mechanism to determine routing and delivery decisions for a given inventory state. Suppose management specifies a value V_i for replenishing one unit of product at customer i. A dispatcher can now evaluate a feasible delivery route as follows. If a set $S = \{s_1, \ldots, s_n\}$ of customers is visited, quantities $d_1, \ldots, d_n$ are delivered, and a cost c_S is incurred. Then the net value of the route equals $\sum_{i \in S} V_i d_i - c_S$. The dispatcher has to choose delivery routes so as to maximize his total net value without stockouts at customers. This mechanism motivates the dispatcher to replenish a customer i whose current inventory level is low, because then d_i can be set large. When faced with the option of expanding the set S of customers to visit on a route

which does not yet use the full vehicle capacity, the dispatcher will consider the incremental cost $c_{S\cup\{k\}} - c_S$ and determine if a quantity d_k can be replenished that is large enough to justify it, i.e., whether $d_k V_k - (c_{S\cup\{k\}} - c_S) > 0$ or $d_k \geqslant (c_{S\cup\{k\}} - C_S)/V_k$.

The key to success in solving management's problem is to set the V_i's in such a way that the dispatcher is motivated to (ideally) minimize the long-run time average replenishment costs. If the dispatcher's total net value is regularly positive, then his performance exceeds management's long range expectations. Management should decrease the V_i's to make them consistent with actual performance. On the other hand, if the dispatcher's total net value is regularly negative, then the V_i's impose unrealistic expectations on the dispatcher and management should increase them. Ideally, management should set the V_i's equal to the lowest achievable marginal costs.

Starting from a dynamic control model of the inventory routing problem, Adelman (2003b) derives the following nonlinear programming relaxation, which computes a long run "average" solution to the inventory routing problem. Let z_R be a decision variable representing the rate at which a subset R of customers is visited together. Furthermore, let $d_{i,R}$ for all $i \in R$ be a decision variable representing the average quantity delivered to customer i on a delivery route visiting subset R. This yields the following formulation:

$$\text{(NLP)} \quad \text{minimize} \quad \sum_{R\subseteq N} C_R z_R \tag{33}$$

subject to

$$\sum_{R\subseteq N} d_{i,R} z_R = u_i, \quad i \in N, \tag{34}$$

$$\sum_{i\in R} d_{i,R} \leqslant Q, \quad R \subseteq N, \tag{35}$$

$$d_{i,R} \leqslant C_i, \quad R \subseteq N, i \in R, \tag{36}$$

$$z_R, d_{i,R} \geqslant 0, \quad R \subseteq N, i \in R. \tag{37}$$

The objective (33) minimizes the long run average replenishment cost. Constraints (34) state that for each customer i the rate at which quantities are replenished must equal the rate at which they are consumed. Constraints (35) state that on average vehicle capacity is satisfied, and constraints (36) state that on average the quantity delivered at customer i is less than the storage capacity. Consider the following linear program

$$\text{(D)} \quad \text{maximize} \quad \sum_{i\in N} u_i V_i \tag{38}$$

subject to

$$\sum_{i\in R} d_{i,R} V_i \leqslant C_R, \quad R \subseteq N, \tag{39}$$

with decision variables V_i. Adelman shows that this semi-infinite linear program is dual to the nonlinear program in that there is no duality gap between them and a version of complementary slackness holds. In (NLP) $d_{i,R}$ is a decision variable while in (D) it is part of the input. The decision variables V_i at optimality are the marginal costs associated with satisfying constraints (34) of (NLP). This means that at optimality $u_i V_i$ is the total allocated cost rate for replenishing customer i in an optimal solution to (NLP). Each V_i can be interpreted as the payment management transfers to the dispatcher for replenishing one unit of product of customer i. Hence, the objective (38) maximizes the total transfer rate, subject to the constraint (39) that the payments can be no larger than the cost of any replenishment. NLP can be solved effectively by means of column generation techniques.

We have opted to focus on only a few research streams with an emphasis on more recent efforts. However, many other researchers have contributed to the inventory routing literature, including Federgruen and Zipkin (1984), Golden et al. (1984), Burns et al. (1985), Larson (1988), Chien et al. (1989), Webb and Larson (1995), Barnes-Schuster and Bassok (1997), Herer and Roundy (1997), Viswanathan and Mathur (1997), Christiansen and Nygreen (1998a, 1998b), Christiansen (1999), Reimann et al. (1999), Waller et al. (1999), Çetinkaya and Lee (2000), Lau et al. (2002), Bertazzi et al. (2002), Savelsbergh and Song (2005), and Song and Savelsbergh (2005).

5 Stochastic vehicle routing problems

Stochastic Vehicle Routing Problems (SVRPs) are extensions of the deterministic VRP in which some components are random. The three most common cases are:

(1) stochastic customers: customer i is present with probability p_i and absent with probability $1 - p_i$;
(2) stochastic demands (to be collected, say): the demand ξ_i of customer i is a random variable;
(3) stochastic times: the service time s_i of customer i and the travel time t_{ij} of edge (i, j) are random variables.

Because some of the data are random it is no longer required to satisfy the constraints for all realizations of the random variables, and new feasibility and optimality concepts are required. With respect to their deterministic counterparts, SVRPs are considerably more difficult to solve. Not only is the notion of a solution different, but some of the properties that were valid in a deterministic context no longer hold in the stochastic case (see, e.g., Dror et al., 1989; Gendreau et al., 1996).

Applications of SVRP arise in a number of settings such as the delivery of meals on wheels (Bartholdi et al., 1983) or of home heating oil (Dror et al., 1985), sludge disposal (Larson, 1988), forklift routing in warehouses

(Bertsimas, 1992), money collection in bank branches (Lambert et al., 1993), and general pickup and delivery operations (Hvattum et al., 2006).

Stochastic VRPs can be formulated and solved in the context of stochastic programming: a first stage or a priori solution is computed, the realizations of the random variables are then disclosed and, in a second stage, a *recourse* or corrective action is applied to the first stage solution. The recourse action usually generates a cost or a saving which may be taken into account when designing the first stage solution. To illustrate, consider a planned vehicle route in an SVRP with stochastic demands. Because demands are stochastic, the vehicle capacity may be attained or exceeded at some customer j before the route is completed. In this case several possible recourse policies are possible. For example, the vehicle could return to the depot to unload and resume collections at customer j (if the vehicle capacity was exceeded at j) or at the successor of j on the route (if the vehicle capacity was attained exactly at j). Another policy would be to plan preventive return trips to the depot in the hope of avoiding higher costs at a later stage (see, e.g., Laporte and Louveaux, 1990; Dror et al., 1993; Yang et al., 2000). A more radical policy would be to re-optimize the route segment following j upon arrival at the depot (see, e.g., Bastian and Rinnooy Kan, 1992; Secomandi, 1998; Haughton, 1998, 2000). The best choice of a recourse policy depends on the time at which information becomes available. For example, information about a customer demand may only be available upon arriving at that customer or when visiting the previous customer, thus allowing for a wider range of recourse actions, such as returning to the depot in anticipation of failure or postponing the visit of a high demand customer. An extensive discussion of recourse policies in the context of availability of information is provided in Dror et al. (1989).

There exist two main solution concepts in stochastic programming. In Chance Constrained Programming (CCP) the first stage problem is solved under the condition that the constraints are satisfied with some probability. For example, one could impose a failure threshold α, i.e., planned vehicle routes should fail with probability at most equal to α. The cost of failure is typically disregarded in this approach. Stewart and Golden (1983) have proposed the first CCP formulation for the VRP with stochastic demands. Using a three-index model they showed that probabilistic constraints could be transformed into a deterministic equivalent form. Laporte et al. (1989) later proposed a similar transformation for a two-index model. The interest of such transformations is that the chance constrained SVRP can then be solved using any of the algorithms available for the deterministic case. In Stochastic Programming with Recourse (SPR) two sets of variables are used: first-stage variables characterize the solution generated before the realization of the random variables, while second-stage variables define the recourse action. The solution cost is defined as the sum of the cost of the first-stage solution and that of the recourse action. The aim of SPR is to design a first-stage solution of least expected total cost.

Stochastic VRPs are usually modeled and solved with the framework of a priori optimization (Bertsimas et al., 1990) or as Markov decision processes (Dror et al., 1989). A priori optimization computes a first-stage solution of least expected cost under a given recourse policy. The most favored a priori optimization methodology is the *integer L-shaped method* (Laporte and Louveaux 1993, 1998) which belongs to the same class as Benders decomposition (Benders, 1962) and the L-shaped method for continuous stochastic programming (Van Slyke and Wets, 1969). While route reoptimization is preferable to a priori optimization from a solution cost point of view, it is computationally more cumbersome. In contrast, a priori optimization entails solving only one instance of an NP-hard problem and produces a more stable and predictable solution (Bertsimas et al., 1990). It is also superior to solving a deterministic VRP instance with expected demands (Louveaux, 1998).

The integer L-shaped method is essentially a variant of branch-and-cut. It operates on a current problem obtained by relaxing integrality requirements and subtour elimination constraints, and by replacing the cost of recourse $Q(x)$ of first-stage solution x by a lower bound θ on its value. Integrality and subtour elimination constraints are gradually satisfied as is commonly done in branch-and-cut algorithms for the deterministic VRP (see, e.g., Naddef and Rinaldi, 2002) while lower bounding functionals on θ, called *optimality cuts*, are introduced into the problem at integer or fractional solutions. The method assumes that a lower bound L on θ is available. In the following description x_{ij} is a binary variable equal to 1 if and only if edge (i, j) is used in the first stage solution.

Step 0. Set the iteration count $\nu := 0$ and introduce the bounding constraint $\theta \geqslant L$ into the current problem. Set the value $\bar{z}$ of the best known solution equal to ∞. At this stage, the only active node corresponds to the initial current problem.

Step 1. Select a pendent node from the list. If none exists stop.

Step 2. Set $\nu := \nu+1$ and solve the current problem. Let (x^ν, θ^ν) be an optimal solution.

Step 3. Check for any subtour elimination constraint violation. If at least one violation can be identified, introduce a suitable number of subtour elimination constraints into the current problem, and return to Step 2. Otherwise, if $cx^\nu + \theta^\nu \geqslant \bar{z}$, fathom the current node and return to Step 1.

Step 4. If the solution is not integer, branch on a fractional variable. Append the corresponding subproblems to the list of pendent nodes and return to Step 1.

Step 5. Compute $Q(x^\nu)$ and set $z^\nu := cx^\nu + Q(x^\nu)$. If $z^\nu < \bar{z}$, set $\bar{z} := z^\nu$.

Step 6. If $\theta^\nu \geqslant Q(x^\nu)$, then fathom the current node and return to Step 1. Otherwise, impose the optimality cut

$$\theta \geqslant L + \big(Q(x^\nu) - L\big)\Bigg(\sum_{1<i<j, x^\nu_{ij}=1} x_{ij} - \sum_{1<i<j} x^\nu_{ij} + 1\Bigg) \tag{40}$$

into the current problem and return to Step 2.

The optimality cut (40) uses the fact that a feasible solution is fully characterized by the x_{ij} variables associated with edges nonincident to the depot. They state that either the current solution must be maintained, in which case the cut becomes $\theta \geqslant Q(x^{\nu})$, or a new solution must be identified, in which case the cut becomes $\theta \geqslant L$ or less and is thus redundant.

Markov decision models are defined on a state space. The system is observed at various transition times corresponding to moments at which a new customer is visited, and new decisions are taken at these moments. The state of the system at a given transition time is described by the set of customers already visited by the vehicle and by its current load. Because the state space is typically very large, this approach can only be applied to relatively small scale instances.

Heuristics for SVRPs are adaptations of methods originally designed for the deterministic case, which can be rather intricate because of the probability computations involved. In particular, computing the expected cost of a vehicle route is itself complicated and it may be advisable to use approximations if such computations are to be performed repeatedly within a search process (see, e.g., Gendreau et al., 1996). In what follows we study some particular classes of SVRPs.

5.1 The vehicle routing problem with stochastic customers

In vehicle routing problems with stochastic customers each vertex i is present with probability p_i. A first-stage solution consists of a set of vehicle routes visiting the depot and each customer exactly once. The set of absent customers is then revealed and the second-stage solution consists of following the first-stage routes while skipping the absent vertices. Jaillet (1985) laid the foundations of this line of research in his study of the Traveling Salesman Problem with Stochastic Customers (TSPSC). He proposed mathematical models and bounds, and he investigated a number of properties of the problem. For example, he showed that the solution of a deterministic TSP can be arbitrarily bad for the TSPSC. Also, even if the TSPSC is defined in a plane with Euclidean distances, an optimal cycle may cross itself, contrary to what happens for the TSP (Flood, 1956). Jézéquel (1985) and Rossi and Gavioli (1987) have proposed a number of heuristics for the TSPSC based on adaptations of the Clarke and Wright (1964) savings principle. Bertsimas (1988) and Bertsimas and Howell (1993) later investigated further properties of the TSPSC and proposed new heuristics, namely methods based on space filling curves (Bartholdi and Platzman, 1982) and on a 2-opt edge interchange mechanism. The first exact algorithm for the TSPSC is an integer L-shaped algorithm developed by Laporte et al. (1994) and capable of solving instances involving up to 50 customers. An extension of the TSPSC, called the Pickup and Delivery Traveling Salesman Problem with Stochastic Customers (PDTSPSC), was recently investigated by Beraldi et al. (2005). In this problem there are n requests, each

consisting of a pickup location and of a delivery location, but request i only materializes with probability p_i. The authors show how to efficiently implement a low complexity interchange heuristic for this problem.

The Vehicle Routing Problem with Stochastic Customers (VRPSC) has been mostly studied in the context of unit demand customers. As in the TSPSC, vehicles follow the first-stage routes while skipping the absent customers and return to the depot to unload when their capacity is reached. This problem was first studied by Jézéquel (1985), Jaillet (1987), and Jaillet and Odoni (1988). The latter reference states two interesting properties of the VRPSC:

(1) even if travel costs are symmetric the overall solution cost is dependent on the direction of travel;
(2) larger vehicle capacities may yield larger solution costs.

Bertsimas' PhD thesis (Bertsimas, 1988) is an excellent source of information on this problem. It describes several properties, bounds and heuristics. Waters (1989) has studied the case of general integer demands and has compared three simple heuristics for this problem.

5.2 The vehicle routing problem with stochastic demands

The Vehicle Routing Problem with Stochastic Demands (VRPSD) has been the most studied stochastic VRP. In this problem customer demands are random and usually (but not always) independent. Tillman (1969) was probably the first to study this problem in a multidepot context. He proposed a savings based heuristic for its solution. The first, major study of the VRPSD can be attributed to Golden and Stewart (1978) who presented a chance constrained model and two recourse models. In the first of these a penalty proportional to vehicle overcapacity is imposed; in the second, the penalty is proportional to the expected demand in excess of the vehicle capacity. Several basic heuristics were implemented and tested. Dror and Trudeau (1986) developed further heuristics and showed that for this problem expected travel cost depends on the direction of travel even in the symmetric case. Again, Bertsimas' thesis (1988) constitutes a major contribution to the study of the VRPSD. It proposes several bounds, asymptotic results and properties for the case where ξ_i is equal to 1 with probability p_i, and equal to 0 otherwise. In their survey paper, Dror et al. (1989) have shown that some properties established by Jaillet (1985, 1988) and Jaillet and Odoni (1988) extend to the VRPSC, namely (1) in an optimal solution a vehicle route may intersect itself; (2) in a Euclidean problem customers are not necessarily visited in the order in which they appear on the convex hull of vertices; (3) segments of an optimal route are not necessarily optimal when considered separately. The latter property can have a major impact on the design of a dynamic algorithm for the VRPSD.

Laporte et al. (1989) proposed a two-index chance constrained model for the VRPSC as well as an associated branch-and-cut algorithm capable of solving instances with $n \leqslant 30$. They also introduced a bounded penalty model in

which the cost of recourse associated with a given route cannot exceed a preset proportion of the first-stage route cost. The best exact solution approach for the VRPSD is again the integer L-shaped algorithm. Séguin (1994) and Gendreau et al. (1995) proposed the first implementation of this method for the solution of the VRPSD and were able to solve instances of up to 70 vertices. The most difficult case arises when the expected filling rate f of the vehicles is large. For example, when $f = 0.3$ instances with $n = 70$ can be solved optimally, but when $f = 1.0$ instances with $n = 10$ can rarely be solved. Using a similar approach, Hjorring and Holt (1999) solved one-vehicle instances ($m = 1$) with $0.95 \leqslant f \leqslant 1.05$ and $n = 90$. Laporte et al. (2002) imposed an additional restriction, namely that the expected demand of a route does not exceed the vehicle capacity, and they also exploited properties of the demand under known distributions (Poisson and normal) in the generation of lower bounding functionals on the cost of recourse. This enabled them to solve larger instances: for Poisson demands they solved instances with $f = 0.9$, $m = 4$, and $n = 25$, or with $m = 2$ and $n = 100$; for normal demands they solved instances with $f = 0.9$, $m = 3$, and $n = 50$.

Dynamic programming was applied by Secomandi (1998) to the VRP with stochastic demands. The largest instance solved to optimality with this method contained only 10 customers. The author also developed Neuro-Dynamic Programming (NDP) algorithms (Secomandi, 1998, 2000, 2003) for the same problem. Neuro-dynamic programming (see, e.g., Bertsekas, 1995) is a heuristic approach used to solve large-scale dynamic programs. It replaces the "cost-to-go" computations by proxies based on simulation and parametric function approximations. Secomandi (2000) compared two NDP implementations for the VRP with stochastic demands: an optimistic approximate iteration policy in which a neural network methodology is used to compute the approximations, and a rollout policy in which the cost-to-go is approximated by means of a heuristic. Computational results show that the second of these two policies is consistently and substantially superior to the first.

5.3 The vehicle routing problem with stochastic customers and demands

The VRP with stochastic customers and demands combines two difficult cases. This problem was first mentioned by Jézéquel (1985), Jaillet (1987), Jaillet and Odoni (1988), and was later formally defined by Bertsimas (1992). A first-stage solution visiting all customers is first constructed, the set of present customers is then revealed and their demand becomes known upon the arrival of the vehicle at the customer's location, routes are followed as planned but absent customers are skipped and the vehicle returns to the depot to unload whenever its capacity becomes attained. Benton and Rossetti (1992) proposed an algorithm which performs route reoptimizations whenever demands are revealed. One major difficulty in solving this problem lies in the computation of the objective function value. Recursions, bounds, asymptotic results, and a comparison of various reoptimization policies are provided by

Bertsimas (1992). Séguin (1994) and Gendreau et al. (1995) developed the first exact algorithm for this problem, based again on the integer L-shaped approach. They solved instances involving up to 46 customers and concluded that stochastic customers are a far more complicating factor than are stochastic demands. In a different study, Gendreau et al. (1996) developed a tabu search algorithm which uses an approximation of the objective function cost in order to ease computations. On a set of 825 instances with $6 \leqslant n \leqslant 46$ for which the optimum was known, an optimal solution was identified in 89.45% of the cases and the average optimality gap was 0.38%.

5.4 The vehicle routing problem with stochastic travel times

In the Vehicle Routing Problem with Stochastic Travel Times (VRPSTT) travel times on the edges and service times at the vertices are random variables. Vehicles follow their planned routes and may incur a penalty if the route duration exceeds a given deadline. It is natural to make this penalty proportional to the elapsed route duration in excess of the deadline (Laporte et al., 1992). Another possibility is to define a penalty proportional to the uncollected demand within the time limit, as is the case in a money collection application studied by Lambert et al. (1993).

The simplest case of the VRPSTT is the Traveling Salesman Problem with Stochastic Travel Time (TSPSTT) in which there is only one vehicle. It was first studied by Leipälä (1978) who computed the expected length of tours with random travel times. A common version of the TSPSTT is the case where the objective is to design a tour having the largest probability of being completed within the deadline. Kao (1978) proposed two heuristics for this problem: one based on dynamic programming and the other on implicit enumeration. Sniedovich (1981) has shown that dynamic programming applied to the same problem can be suboptimal because the monotonicity property required for this method is not satisfied in the TSPSTT. Carraway et al. (1989) later developed a so-called generalized dynamic programming algorithm that overcomes this difficulty. Kenyon and Morton (2003) have shown that an optimal TSPSTT can be identified by solving a deterministic TSP in which the travel and service times are replaced by their mean values. Verweij et al. (2003) have developed a heuristic for the case where a penalty proportional to route duration in excess of the deadline is incurred. The method uses a sample average approximation technique in which a sample of instance realizations is drawn and each is solved optimally by means of a deterministic technique. By repeating the method with different samples a statistical estimate of the optimality gap can be computed.

Laporte et al. (1992) were probably the first to provide exact algorithms for the VRPSTT. They formulated the chance constrained version of the problem, and they modeled a recourse version of the problem in which a penalty proportional to route duration in excess of the deadline is incurred. The problem was solved optimally by means of an integer L-shaped algorithm for $10 \leqslant n \leqslant 20$

and two to five travel time scenarios (each scenario corresponds to a different travel speed for the entire network). In a more recent study, Kenyon and Morton (2003) have investigated properties of VRPSTT solutions and have developed bounds on the objective function value. They have developed a heuristic that combines branch-and-cut and Monte Carlo simulation which, if run to completion, terminates with a solution value within a preset percentage of the optimum.

Finally, vehicle routing with stochastic travel time is frequently encountered in pickup and delivery problems such as those arising in truckload operations. Wang and Regan (2001) have proposed models for this class of problems under the presence of time windows.

Acknowledgements

This work has been supported by the Canadian Natural Sciences and Engineering Research Council under Grants 227837-00 and OGP0039682, by the Ministero dell'Università e della Ricerca (MIUR), and by the Consiglio Nazionale delle Ricerche (CNR), Italy. This support is gratefully acknowledged.

References

Achuthan, N.R., Caccetta, L., Hill, S.P. (1996). A new subtour elimination constraint for the vehicle routing problem. *European Journal of Operational Research* 91, 573–586.

Achuthan, N.R., Caccetta, L., Hill, S.P. (2003). An improved branch and cut algorithm for the capacitated vehicle routing problem. *Transportation Science* 37, 153–169.

Adelman, D. (2003a). Price-directed replenishment of subsets: Methodology and its application to inventory routing. *Manufacturing & Service Operations Management* 5, 348–371.

Adelman, D. (2003b). Internal transfer pricing for a decentralized operation with a shared supplier. Working paper, Graduate School of Business, The University of Chicago.

Adelman, D. (2004). A price-directed approach to stochastic inventory/routing. *Operations Research* 52, 499–514.

Agarwal, Y., Mathur, K., Salkin, H.M. (1989). A set-partitioning-based exact algorithm for the vehicle routing problem. *Networks* 19, 731–749.

Altinkemer, K., Gavish, B. (1991). Parallel savings based heuristic for the delivery problem. *Operations Research* 39, 456–469.

Anily, S., Federgruen, A. (1990). One warehouse multiple retailer systems with vehicle routing costs. *Management Science* 36, 92–114.

Anily, S., Federgruen, A. (1991). Rejoinder to "One warehouse multiple retailer systems with vehicle routing costs". *Management Science* 37, 1497–1499.

Anily, S., Federgruen, A. (1993). Two-echelon distribution systems with vehicle routing costs and central inventories. *Operations Research* 41, 37–47.

Araque, J.R., Hall, L., Magnanti, T.L. (1990). Capacitated trees, capacitated routing and associated polyhedra. Discussion Paper 90-61, CORE, University of Louvain-la-Neuve, Belgium.

Araque, J.R., Kudva, G., Morin, T., Pekny, J.F. (1994). A branch-and-cut algorithm for vehicle routing problems. *Annals of Operations Research* 50, 37–59.

Augerat, P., Belenguer, J.M., Benavent, E., Corberán, A., Naddef, D., Rinaldi, G. (1995). Computational results with a branch and cut code for the capacitated vehicle routing problem. Technical Report RR 949-M, Université Joseph Fourier, Grenoble.

Augerat, P., Belenguer, J.M., Benavent, E., Corberán, A., Naddef, D. (1999). Separating capacity inequalities in the CVRP using tabu search. *European Journal of Operational Research* 106, 546–557.

Badeau, P., Gendreau, M., Guertin, F., Potvin, J.-Y., Taillard, É.D. (1997). A parallel tabu search heuristic for the vehicle routing problem with time windows. *Transportation Research C* 5, 109–122.

Baker, E., Schaffer, J. (1986). Computational experience with branch exchange heuristics for vehicle routing problems with time window constraints. *American Journal of Mathematical and Management Sciences* 6, 261–300.

Baldacci, R., Hadjiconstantinou, E., Mingozzi, A. (2004). An exact algorithm for the capacitated vehicle routing problem based on a two-commodity network flow formulation. *Operations Research* 52, 723–738.

Baldacci, R., Bodin, L., Mingozzi, A. (2006). The multiple disposal facilities and multiple inventory locations rollon–rolloff vehicle routing problem. *Computers & Operations Research* 33, 2667–2702.

Balinski, M., Quandt, R. (1964). On an integer program for a delivery problem. *Operations Research* 12, 300–304.

Bard, J.F., Huang, L., Dror, M., Jaillet, P. (1998a). A branch and cut algorithm for the VRP with satellite facilities. *IIE Transactions* 30, 831–834.

Bard, J.F., Huang, L., Jaillet, P., Dror, M. (1998b). A decomposition approach to the inventory routing problem with satellite facilities. *Transportation Science* 32, 189–203.

Bard, J.F., Kontoravdis, G., Yu, G. (2002). A branch-and-cut procedure for the vehicle routing problem with time windows. *Transportation Science* 36, 250–269.

Barnes-Schuster, D., Bassok, Y. (1997). Direct shipping and the dynamic single-depot/multi-retailer inventory system. *European Journal of Operational Research* 101, 509–518.

Bartholdi, J.J., Platzman, L.K. (1982). An $N \log N$ planar traveling salesman heuristic based on space-filling curves. *Operations Research Letters* 1, 121–125.

Bartholdi, J.J., Platzman, L.K., Collins, R.L., Warden, W.H. (1983). A minimal technology routing system for meals on wheels. *Interfaces* 13 (3), 1–8.

Bastian, C., Rinnooy Kan, A.H.G. (1992). The stochastic vehicle routing problem revisited. *European Journal of Operational Research* 56, 407–412.

Battiti, R., Tecchiolli, G. (1994). The reactive tabu search. *ORSA Journal on Computing* 6, 126–140.

Bean, J.C. (1994). Genetic algorithms and random keys for the sequencing and optimization. *ORSA Journal on Computing* 6, 154–160.

Beasley, J.E. (1983). Route-first cluster-second methods for vehicle routing. *Omega* 11, 403–408.

Bell, W., Dalberto, L., Fisher, M.L., Greenfield, A., Jaikumar, R., Kedia, P., Mack, R., Prutzman, P. (1983). Improving the distribution of industrial gases with an on-line computerized routing and scheduling optimizer. *Interfaces* 13 (6), 4–23.

Benders, J.F. (1962). Partitioning procedures for solving mixed variables programming problems. *Numerische Mathematik* 4, 238–252.

Bent, R., Van Hentenryck, P. (2004). A two-stage hybrid local search for the vehicle routing problem with time windows. *Transportation Science* 38, 515–530.

Benton, W.C., Rossetti, M.D. (1992). The vehicle scheduling problem with intermittent customer demands. *Computers & Operations Research* 19, 521–531.

Beraldi, P., Ghiani, G., Laporte, G., Musmanno, G. (2005). Efficient neighbourhood search for the probabilistic pickup and delivery travelling salesman problem. *Networks* 46, 195–198.

Berger, J., Barkaoui, M. (2004). A new hybrid genetic algorithm for the capacitated vehicle routing problem. *Journal of the Operational Research Society* 54, 1254–1262.

Berger, J., Barkaoui, M., Bräysy, O. (2003). A route-directed hybrid genetic approach for the vehicle routing problem with time windows. *INFOR* 41, 179–194.

Bertazzi, L., Paletta, G., Speranza, M.G. (2002). Deterministic order-up-to level policies in an inventory routing problem. *Transportation Science* 36, 119–132.

Bertsekas, D.P. (1995). *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA.

Bertsimas, D.J. (1988). Probabilistic combinatorial optimization problems. PhD thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.
Bertsimas, D.J. (1992). A vehicle routing problem with stochastic demand. *Operations Research* 40, 574–585.
Bertsimas, D.J., Howell, L.H. (1993). Further results on the probabilistic traveling salesman problem. *European Journal of Operational Research* 65, 68–95.
Bertsimas, D.J., Simchi-Levi, D. (1996). A new generation of vehicle routing research: Robust algorithms addressing uncertainty. *Operations Research* 44, 286–304.
Bertsimas, D.J., Jaillet, P., Odoni, A.R. (1990). A priori optimisation. *Operations Research* 38, 1019–1033.
Blasum, U., Hochstättler, W. (2000). Application of the branch and cut method to the vehicle routing problem. Technical Report ZPR2000-386, ZPR, Universität zu Köln. Available at *http://www.zaik.uni-koeln.de/~paper*.
Bozkaya, B., Erkut, E., Laporte, G. (2003). A tabu search algorithm and adaptive memory procedure for political districting. *European Journal of Operational Research* 144, 12–26.
Bramel, J., Simchi-Levi, D. (1995). A location based heuristic for general routing problems. *Operations Research* 43, 649–660.
Bramel, J., Simchi-Levi, D. (1997). On the effectiveness of set covering formulations for the vehicle routing problem with time windows. *Operations Research* 45, 295–301.
Bramel, J., Simchi-Levi, D. (2002). Set-covering-based algorithms for the capacitated VRP. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 85–108.
Brandão, J. (1998). Metaheuristic for the vehicle routing problem with time windows. In: Voss, S., Martello, S., Osman, I.H., Roucairol, C. (Eds.), *Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization*. Kluwer Academic, Boston, pp. 19–36.
Bräysy, O. (2002). Fast local searches for the vehicle routing problem with time windows. *INFOR* 40, 319–330.
Bräysy, O. (2003). A reactive variable neighborhood search for the vehicle routing problem with time windows. *INFORMS Journal on Computing* 15, 347–368.
Bräysy, O., Gendreau, M. (2005a). Vehicle routing problem with time windows, Part I: Route construction and local search algorithms. *Transportation Science* 39, 104–118.
Bräysy, O., Gendreau, M. (2005b). Vehicle routing problem with time windows, Part II: Metaheuristics. *Transportation Science* 39, 119–139.
Burns, L.D., Hall, R.W., Blumenfeld, D.E., Daganzo, C.F. (1985). Distribution strategies that minimize transportation and inventory costs. *Operations Research* 33, 469–490.
Campbell, A.M., Savelsbergh, M.W.P. (2004a). Delivery volume optimization. *Transportation Science* 38, 210–223.
Campbell, A.M., Savelsbergh, M.W.P. (2004b). A decomposition approach for the inventory-routing problem. *Transportation Science* 38, 488–502.
Campbell, A.M., Savelsbergh, M.W.P. (2004c). Efficient insertion heuristics for vehicle routing and scheduling problems. *Transportation Science* 38, 269–378.
Campbell, A.M., Clarke, L., Kleywegt, A.J., Savelsbergh, M.W.P. (1998). The inventory routing problem. In: Crainic, T.G., Laporte, G. (Eds.), *Fleet Management and Logistics*. Kluwer Academic, Boston, pp. 95–112.
Campos, V., Corberán, A., Mota, E. (1991). Polyhedral results for a vehicle routing problem. *European Journal of Operational Research* 52, 75–85.
Carraway, R.L., Morin, T.L., Moskowitz, H. (1989). Generalized dynamic programming for stochastic combinatorial optimization. *Operations Research* 37, 819–829.
Çetinkaya, S., Lee, C.Y. (2000). Stock replenishment and shipment scheduling for vendor managed inventory systems. *Management Science* 46, 217–232.
Chabrier, A. (2006). Vehicle routing problem with elementary shortest path based column generation. *Computers & Operations Research* 33, 2972–2990.
Chan, L.M., Federgruen, A., Simchi-Levi, D. (1998). Probabilistic analyses and practical algorithms for inventory-routing models. *Operations Research* 46, 96–106.

Chiang, W.-C., Russell, R.A. (1997). A reactive tabu search metaheuristic for the vehicle routing problem with time windows. *INFORMS Journal on Computing* 9, 417–430.
Chien, T., Balakrishnan, A., Wong, R. (1989). An integrated inventory allocation and vehicle routing problem. *Transportation Science* 23, 67–76.
Christiansen, M. (1999). Decomposition of a combined inventory and time constrained ship routing problem. *Transportation Science* 33, 3–16.
Christiansen, M., Nygreen, B. (1998a). A method for solving ship routing problems with inventory constraints. *Annals of Operations Research* 81, 357–378.
Christiansen, M., Nygreen, B. (1998b). Modelling path flows for a combined ship routing and inventory management problem. *Annals of Operations Research* 82, 391–412.
Christofides, N., Mingozzi, A. (1989). Vehicle routing: Practical and algorithmic aspects. In: Van Rijn, C.F.H. (Ed.), *Logistics: Where Ends Have to Meet*. Pergamon, Oxford, pp. 30–48.
Christofides, N., Mingozzi, A., Toth, P. (1979). The vehicle routing problem. In: Christofides, N., Mingozzi, A., Toth, P., Sandi, C. (Eds.), *Combinatorial Optimization*. Wiley, Chichester, pp. 315–338.
Christofides, N., Mingozzi, A., Toth, P. (1981a). Exact algorithms for the vehicle routing problem based on the spanning tree and shortest path relaxations. *Mathematical Programming* 20, 255–282.
Christofides, N., Mingozzi, A., Toth, P. (1981b). State-space relaxation procedures for the computation of bounds to routing problems. *Networks* 11, 145–164.
Clarke, G., Wright, J.W. (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations Research* 12, 568–581.
Cook, W., Rich, J.L. (1999). A parallel cutting-plane algorithm for the vehicle routing problem with time windows. Technical Report TR99-04, Computational and Applied Mathematics Department, Rice University, TX.
Cordeau, J.-F., Laporte, G. (2001). A tabu search algorithm for the site dependent vehicle routing problem with time windows. *INFOR* 39, 292–298.
Cordeau, J.-F., Laporte, G. (2004). Tabu search heuristics for the vehicle routing problem. In: Rego, C., Alidaee, B. (Eds.), *Metaheuristic Optimization via Memory and Evolution: Tabu Search and Scatter Search*. Kluwer Academic, Boston, pp. 145–163.
Cordeau, J.-F., Gendreau, M., Laporte, G. (1997). A tabu search heuristic for periodic and multi-depot vehicle routing problems. *Networks* 30, 105–119.
Cordeau, J.-F., Laporte, G., Mercier, A. (2001). A unified tabu search heuristic for vehicle routing problems with time windows. *Journal of the Operational Research Society* 52, 928–936.
Cordeau, J.-F., Desaulniers, G., Desrosiers, J., Solomon, M.M., Soumis, F. (2002a). VRP with Time Windows. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 157–193.
Cordeau, J.-F., Gendreau, M., Laporte, G., Potvin, J.-Y., Semet, F. (2002b). A guide to vehicle routing heuristics. *Journal of the Operational Research Society* 53, 512–522.
Cordeau, J.-F., Laporte, G., Mercier, A. (2004). An improved tabu search algorithm for the handling of route duration constraints in vehicle routing problems with time windows. *Journal of the Operational Research Society* 55, 542–546.
Cordeau, J.-F., Gendreau, M., Hertz, A., Laporte, G., Sormany, J.-S. (2005). New heuristics for the vehicle routing problem. In: Langevin, A., Riopel, D. (Eds.), *Logistics Systems: Design and Optimization*. Springer-Verlag, New York, pp. 279–297.
Cordone, R., Wolfler Calvo, R. (2001). A heuristic for the vehicle routing problem with time windows. *Journal of Heuristics* 7, 107–129.
Cornuéjols, G., Harche, F. (1993). Polyhedral study of the capacitated vehicle routing problem. *Mathematical Programming* 60, 21–52.
Croes, A. (1958). A method for solving traveling salesman problems. *Operations Research* 6, 791–812.
Danna, E., Le Pape, C. (2003). Accelerating branch-and-price with local search: A case study on the vehicle routing problem with time windows. Technical Report 03-006, ILOG.
Dantzig, G.B., Ramser, J.M. (1959). The truck dispatching problem. *Management Science* 6, 81–91.
Dantzig, G.B., Wolfe, P. (1960). Decomposition principle for linear programming. *Operations Research* 8, 101–111.

De Backer, B., Furnon, V., Kilby, P., Prosser, P., Shaw, P. (2000). Solving vehicle routing problems using constraint programming and metaheuristics. *Journal of Heuristics* 6, 501–523.
Dell'Amico, M., Toth, P. (2000). Algorithms and codes for dense assignment problems: The state of the art. *Discrete Applied Mathematics* 100, 17–48.
Desrochers, M., Laporte, G. (1991). Improvements and extensions to the Miller–Tucker–Zemlin subtour elimination constraints. *Operations Research Letters* 10, 27–36.
Desrochers, M., Soumis, F. (1988). A generalized permanent labeling algorithm for the shortest path problem with time windows. *INFOR* 26, 191–212.
Desrochers, M., Verhoog, T.W. (1989). A matching based savings algorithm for the vehicle routing problem. Les Cahiers du GERAD G–89–04, HEC Montréal.
Desrochers, M., Lenstra, J.K., Savelsbergh, M.W.P., Soumis, F. (1988). Vehicle routing with time windows: Optimization and approximation. In: Golden, B.L., Assad, A.A. (Eds.), *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, pp. 65–84.
Desrochers, M., Desrosiers, J., Solomon, M.M. (1992). A new optimization algorithm for the vehicle routing problem with time windows. *Operations Research* 40, 342–354.
Dorigo, M., Di Caro, G., Gambardella, L.M. (1999). Ant algorithms for discrete optimization. *Artificial Life* 5, 137–172.
Drezner, Z. (2003). A new genetic algorithm for the quadratic assignment problem. *INFORMS Journal on Computing* 15, 320–330.
Dror, M., Ball, M.O. (1987). Inventory/routing: Reduction from an annual to a short period problem. *Naval Research Logistics Quarterly* 34, 891–905.
Dror, M., Trudeau, P. (1986). Stochastic vehicle routing with modified savings algorithm. *European Journal of Operational Reserach* 23, 228–235.
Dror, M., Ball, M.O., Golden, B.L. (1985). A computational comparison of algorithms for the inventory routing problem. *Annals of Operations Research* 4, 3–23.
Dror, M., Laporte, G., Trudeau, P. (1989). Vehicle routing with stochastic demands: Properties and solution frameworks. *Transportation Science* 23, 166–176.
Dror, M., Laporte, G., Louveaux, F.V. (1993). Vehicle routing with stochastic demands and restricted failures. *Zeitschrift für Operations Research* 37, 273–283.
Dueck, G. (1990). New optimization heuristics, the great deluge algorithm and the record-to-record travel. Technical report, IBM Germany, Heidelberg Scientific Center.
Dueck, G. (1993). New optimization heuristics: The great deluge algorithm and the record-to-record travel. *Journal of Computational Physics* 104, 86–92.
Ergun, Ö., Orlin, J.B., Steele-Feldman, A. (2003). Creating very large scale neighborhoods out of smaller ones by compounding moves: A study on the vehicle routing problem. MIT Sloan working Paper 4393-02, Massachusetts Institute of Technology, Cambridge, MA.
Federgruen, A., Zipkin, P. (1984). A combined vehicle routing and inventory allocation problem. *Operations Research* 32, 1019–1036.
Fischetti, M., Toth, P. (1989). An additive bounding procedure for combinatorial optimization problems. *Operations Research* 37, 319–328.
Fischetti, M., Toth, P., Vigo, D. (1994). A branch-and-bound algorithm for the capacitated vehicle routing problem on directed graphs. *Operations Research* 42, 846–859.
Fisher, M.L. (1994). Optimal solution of vehicle routing problems using minimum k-trees. *Operations Research* 42, 626–642.
Fisher, M.L., Jaikumar, R. (1981). A generalized assignment heuristic for the vehicle routing problem. *Networks* 11, 109–124.
Fisher, M.L., Greenfield, A., Jaikumar, R., Kedia, P. (1982). Real-time scheduling of a bulk delivery fleet: Practical application of Lagrangean relaxation. Technical report, The Wharton School, University of Pennsylvania.
Fisher, M.L., Jörnsten, K.O., Madsen, O.B.G. (1997). Vehicle routing with time windows – two optimization algorithms. *Operations Research* 45, 488–492.
Flood, M.M. (1956). The travelling salesman problem. *Operations Research* 4, 61–75.
Foster, B.A., Ryan, D.M. (1976). An integer programming approach to the vehicle scheduling problem. *Operations Research* 27, 367–384.

Fukasawa, R., Longo, H., Lysgaard, J., Poggi de Aragão, M., Reis, M., Uchoa, E., Werneck, R.F. (2006). Robust branch-and-cut-and-price for the capacitated vehicle routing problem. *Mathematical Programming A* 106, 491–511.

Gallego, G., Simchi-Levi, D. (1990). On the effectiveness of direct shipping strategy for the one-warehouse multi-retailer *r*-systems. *Management Science* 36, 240–243.

Gambardella, L.M., Taillard, É.D., Agazzi, G. (1999). MACS-VRPTW: A multiple ant colony system for vehicle routing problems with time windows. In: Corne, D., Dorigo, M., Glover, F. (Eds.), *New Ideas in Optimization*. McGraw-Hill, London, pp. 63–76.

Gaur, V., Fisher, M.L. (2004). A periodic inventory routing problem at a supermarket chain. *Operations Research* 52, 813–822.

Gehring, H., Homberger, J. (2002). Parallelization of a two-phase metaheuristic for routing problems with time windows. *Journal of Heuristics* 8, 251–276.

Gendreau, M., Hertz, A., Laporte, G. (1992). New insertion and post-optimization procedures for the traveling selesman problem. *Operations Research* 40, 1083–1094.

Gendreau, M., Hertz, A., Laporte, G. (1994). A tabu search heuristic for the vehicle routing problem. *Management Science* 40, 1276–1290.

Gendreau, M., Laporte, G., Séguin, R. (1995). An exact algorithm for the vehicle routing problem with stochastic customers and demands. *Transportation Science* 29, 143–155.

Gendreau, M., Laporte, G., Séguin, R. (1996). A tabu search algorithm for the vehicle routing problem with stochastic demands and customers. *Operations Research* 44, 469–477.

Gendreau, M., Hertz, A., Laporte, G., Stan, M. (1998). A generalized insertion heuristic for the traveling salesman problem with time windows. *Operations Research* 43, 330–335.

Gendreau, M., Laporte, G., Potvin, J.-Y. (2002). Metaheuristics for the capacitated VRP. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 129–154.

Ghaziri, H. (1993). Algorithmes connexionistes pour l'optimisation combinatoire. Thèse de doctorat, École Polytechnique Fédérale de Lausanne, Switzerland.

Ghiani, G., Laporte, G., Semet, F. (2006). The black and white traveling salesman problem. *Operations Research* 54, 366–378.

Gillett, B.E., Miller, L.R. (1974). A heuristic algorithm for the vehicle-dispatch problem. *Operations Research* 21, 340–349.

Glover, F. (1992). New ejection chain and alternating path methods for traveling salesman problems. In: Balci, O., Sharda, R., Zenios, S. (Eds.), *Computer Science and Operations Research: New Developments in Their Interfaces*. Pergamon, Oxford, pp. 449–509.

Golden, B.L., Stewart, W.R. (1978). Vehicle routing with probabilistic demands. In: Hogben, D., Fife, D. (Eds.) *Computer Science and Statistics: Tenth Annual Symposium on the Interface. NBS Special Publication*, vol. 503, pp. 252–259.

Golden, B.L., Magnanti, T.L., Nguyen, H.Q. (1977). Implementing vehicle routing algorithms. *Networks* 7, 113–148.

Golden, B.L., Assad, A.A., Dahl, R. (1984). Analysis of a large scale vehicle routing problem with an inventory component. *Large Scale Systems* 7, 181–190.

Golden, B.L., Wasil, E.A., Kelly, J.P., Chao, I-M. (1998). Metaheuristics in vehicle routing. In: Crainic, T.G., Laporte, G. (Eds.), *Fleet Management and Logistics*. Kluwer Academic, Boston, pp. 33–56.

Golden, B.L., Assad, A.A., Wasil, E.A. (2002). Routing vehicles in the real world: Applications in the solid waste, beverage, food, dairy, and newspaper industries. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 245–286.

Gouveia, L. (1995). A result on projection for the vehicle routing problem. *Journal of Operational Research Society* 85, 610–624.

Hadjiconstantinou, E., Christofides, N., Mingozzi, A. (1995). A new exact algorithm for the vehicle routing problem based on *q*-paths and *k*-shortest paths relaxations. *Annals of Operations Research* 61, 21–43.

Haimovich, M., Rinnooy Kan, A.H.G. (1985). Bounds and heuristics for capacitated routing problems. *Mathematics of Operations Research* 10, 527–542.

Haughton, M.A. (1998). The performance of route modification and demand stabilization strategies in stochastic vehicle routing. *Transportation Research* 32, 551–566.

Haughton, M.A. (2000). Quantifying the benefits of route reoptimisation under stochastic customer demands. *Journal of the Operational Research Society* 51, 320–332.

Held, M., Karp, R.M. (1971). The traveling salesman problem and minimum spanning trees: Part II. *Mathematical Programming* 1, 6–25.

Held, M., Wolfe, P., Crowder, M.P. (1974). Validation of the subgradient optimization. *Mathematical Programming* 6, 62–88.

Herer, Y., Roundy, R. (1997). Heuristics for a one-warehouse multiretailer distribution problem with performance bounds. *Operations Research* 45, 102–115.

Hjorring, C., Holt, J. (1999). New optimality cuts for a single-vehicle stochastic routing problem. *Annals of Operations Research* 86, 569–585.

Homberger, J., Gehring, H. (1999). Two evolutionary metaheuristics for the vehicle routing problem with time windows. *INFOR* 37, 297–318.

Houck, D.J., Picard, J.-C., Queyranne, M., Vemuganti, R.R. (1980). The traveling salesman problem as a constrained shortest path problem: Theory and computational experience. *Opsearch* 17, 93–109.

Hvattum, L.M., Løkketangen, A., Laporte, G. (2006). Solving a dynamic and stochastic vehicle routing problem with a sample scenario hedging heuristic. *Transportation Science*, in press.

Ioannou, G., Kritikos, M., Prastacos, G. (2001). A greedy look-ahead heuristic for the vehicle routing problem with time windows. *Journal of the Operational Research Society* 52, 523–537.

Irnich, S., Villeneuve, D. (2003). The shortest path problem with resource constraints and k-cycle elimination for $k \geqslant 3$. Technical report, Rheinisch-Westfälische Technische Hochschule, Aachen, Germany.

Jaillet, P. (1985). Probabilistic traveling salesman problem. PhD thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Jaillet, P. (1987). Stochastic routing problems. In: Andreatta, G., Mason, F., Serafini, P. (Eds.), *Stochastics in Combinatorial Optimization*. World Scientific, Singapore, pp. 197–213.

Jaillet, P. (1988). A priori solution of a traveling salesman problem in which a random subset of the customers are visited. *Operations Research* 36, 929–936.

Jaillet, P., Odoni, A.R. (1988). The probabilistic vehicle routing problem. In: Golden, B.L., Assad, A.A. (Eds.), *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, pp. 293–318.

Jaillet, P., Bard, J.F., Huang, L., Dror, M. (2002). Delivery cost approximations for inventory routing problems in a rolling horizon framework. *Transportation Science* 3, 292–300.

Jézéquel, A. (1985). Probabilistic vehicle routing problems. MSc dissertation, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Kallehauge, B., Larsen, J., Madsen, O.B.G. (2006). Lagrangean duality applied to the vehicle routing with time windows. *Computers & Operations Research* 33, 1464–1487.

Kao, E.P.C. (1978). A preference order dynamic program for a stochastic traveling salesman problem. *Operations Research* 26, 1033–1045.

Kenyon, A.S., Morton, D.P. (2003). Stochastic vehicle routing with random travel times. *Transportation Science* 37, 69–82.

Kilby, P.J., Prosser, P., Shaw, P. (1998). Guided local search for the vehicle routing problem with time windows. In: Voss, S., Martello, S., Osman, I.H., Roucairol, C. (Eds.), *Meta Heuristics: Advances and Trends in Local Search Paradigms for Optimisation*. Kluwer Academic, Boston, pp. 473–486.

Kindervater, G.A.P., Savelsbergh, M.W.P. (1997). Vehicle routing: Handling edge exchanges. In: Aarts, E.H.L., Lenstra, J.K. (Eds.), *Local Search in Combinatorial Optimization*. Wiley, Chichester, pp. 337–360.

Kleywegt, A.J., Nori, V., Savelsbergh, M.W.L. (2002). The stochastic inventory routing problem with direct deliveries. *Transportation Science* 36, 94–118.

Kleywegt, A.J., Nori, V., Savelsbergh, M.W.L. (2004). Dynamic programming approximations for a stochastic inventory routing problem. *Transportation Science* 38, 42–70.

Kohl, N., Madsen, O.B.G. (1997). An optimization algorithm for the vehicle routing problem with time windows based on Lagrangean relaxation. *Operations Research* 45, 395–406.

Kohl, N., Desrosiers, J., Madsen, O.B.G., Solomon, M.M., Soumis, F. (1999). 2-path cuts for the vehicle routing problem with time windows. *Transportation Science* 33, 101–116.

Kolen, A.W.J., Rinnooy Kan, A.H.G., Trienekens, H.W.J.M. (1987). Vehicle routing with time windows. *Operations Research* 35, 256–273.

Kontoravdis, G., Bard, J.F. (1995). A GRASP for the vehicle routing problem with time windows. *ORSA Journal on Computing* 7, 10–23.

Lambert, V., Laporte, G., Louveaux, F.V. (1993). Designing collection routes through bank branches. *Computers & Operations Research* 20, 783–791.

Laporte, G., Louveaux, F.V. (1990). Formulations and bounds for the stochastic capacitated vehicle routing problem with uncertain supplies. In: Gabzewicz, J., Richard, J.-F., Wolsey, L. (Eds.), *Economic Decision Making: Games, Econometrics and Optimisation*. North-Holland, Amsterdam, pp. 443–455.

Laporte, G., Louveaux, F.V. (1993). The integer *L*-shaped method for stochastic integer programs with complete recourse. *Operations Research Letters* 13, 133–142.

Laporte, G., Louveaux, F.V. (1998). Solving stochastic routing problems with the integer *L*-shaped method. In: Crainic, T.G., Laporte, G. (Eds.), *Fleet Management and Logistics*. Kluwer Academic, Boston, pp. 159–167.

Laporte, G., Nobert, Y. (1987). Exact algorithms for the vehicle routing problem. *Annals of Discrete Mathematics* 31, 147–184.

Laporte, G., Semet, F. (2002). Classical heuristics for the capacitated VRP. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 109–128.

Laporte, G., Nobert, Y., Desrochers, M. (1985). Optimal routing under capacity and distance restrictions. *Operations Research* 33, 1050–1073.

Laporte, G., Mercure, H., Nobert, Y. (1986). An exact algorithm for the asymmetrical capacitated vehicle routing problem. *Networks* 16, 33–46.

Laporte, G., Louveaux, F.V., Mercure, H. (1989). Models and exact solutions for a class of stochastic location-routing problems. *European Journal of Operational Research* 39, 71–78.

Laporte, G., Louveaux, F.V., Mercure, H. (1992). The vehicle routing problem with stochastic travel times. *Transportation Science* 26, 161–170.

Laporte, G., Louveaux, F.V., Mercure, H. (1994). A priori optimization of the probabilistic traveling salesman problem. *Operations Research* 42, 543–549.

Laporte, G., Louveaux, F.V., Van hamme, L. (2002). An integer *L*-shaped algorithm for the capacitated vehicle routing problem with stochastic demands. *Operations Research* 50, 415–423.

Larson, R.C. (1988). Transportation of sludge to the 106-mile site: An inventory routing problem for fleet sizing and logistic system design. *Transportation Science* 22, 186–198.

Lau, H.C., Liu, Q., Ono, H. (2002). Integrating local search and network flow to solve the inventory routing problem. *American Association for Artificial Intelligence* 2, 9–14.

Lau, H.C., Sim, M., Teo, K.M. (2003). Vehicle routing problem with time windows and a limited number of vehicles. *European Journal of Operational Research* 148, 559–569.

Leipälä, T. (1978). On the solutions of stochastic traveling salesman problems. *European Journal of Operational Research* 2, 291–297.

Letchford, A., Eglese, R.W., Lysgaard, J. (2002). Multistars, partial multistars and the capacitated vehicle routing problem. *Mathematical Programming* 94, 21–40.

Li, F., Golden, B.L., Wasil, E.A. (2005). Very large-scale vehicle routing: New test problems, algorithms and results. *Computers & Operations Research* 32, 1165–1179.

Li, H., Lim, A. (2003). Local search with annealing-like restarts to solve the VRPTW. *European Journal of Operational Research* 150, 115–127.

Lin, S. (1965). Computer solutions of the travelling salesman problem. *Bell System Technical Journal* 44, 2245–2269.

Louveaux, F.V. (1998). An introduction to stochastic transportation models. In: Labbé, M., Laporte, G., Tanczos, K., Toint, P. (Eds.), *Operations Research and Decision Aid Methodologies in Traffic and Transportation Management*. *NATO ASI Series F: Computer and Systems Sciences*, vol. 166. Springer-Verlag, Berlin/Heidelberg, pp. 244–263.

Lysgaard, J., Letchford, A., Eglese, R.W. (2004). A new branch-and-cut algorithm for the capacitated vehicle routing problem. *Mathematical Programming* 100, 423–445.

Martinhon, C., Lucena, A., Maculan, N. (2000). A relax and cut algorithm for the vehicle routing problem. Technical Report RT-05/00, Universidade Federal Fluminense, Niterói, Brasil.

Mester, D., Bräysy, O. (2005). Active guided evolution strategies for large scale vehicle routing problem with time windows. *Computers & Operations Research* 32, 1593–1614.

Miller, D.L. (1995). A matching based exact algorithm for capacitated vehicle routing problems. *ORSA Journal on Computing* 7, 1–9.

Miller, D.L., Pekny, J.F. (1995). A staged primal-dual algorithm for perfect *b*-matching with edge capacities. *ORSA Journal on Computing* 7, 298–320.

Mingozzi, A., Christofides, N., Hadjiconstantinou, E. (1994). An exact algorithm for the vehicle routing problem based on the set partitioning formulation. Technical report, Department of Mathematics, University of Bologna, Italy.

Minkoff, A. (1993). A Markov decision model and decomposition heuristic for dynamic vehicle dispatching. *Operations Research* 41, 77–90.

Mladenović, N., Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research* 24, 1097–1100.

Mole, R.H., Jameson, S.R. (1976). A sequential route-building algorithm employing a generalized savings criterion. *Operational Research Quarterly* 27, 503–511.

Moscato, P., Cotta, C. (2003). A gentle introduction to memetic algorithms. In: Glover, F., Kochenberger, G.A. (Eds.), *Handbook of Metaheuristics*. Kluwer Academic, Boston, pp. 105–144.

Naddef, D., Rinaldi, G. (2002). Branch-and-cut algorithms for the capacitated VRP. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 53–84.

Nelson, M.D., Nygard, K.E., Griffin, J.H., Shreve, W.E. (1985). Implementation techniques for the vehicle routing problem. *Computers & Operations Research* 12, 273–283.

Or, I. (1976). Traveling salesman-type combinatorial problems and their relation to the logistics of blood banking. PhD thesis, Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL.

Osman, I.H. (1993). Metastrategy simulated annealing and tabu search algorithms for the vehicle routing problem. *Annals of Operations Research* 41, 421–451.

Paessens, H. (1988). The savings algorithm for the vehicle routing problem. *European Journal of Operational Research* 34, 336–344.

Potvin, J.-Y. (1996). Genetic algorithms for the traveling salesman problem. *Annals of Operations Research* 63, 339–370.

Potvin, J.-Y., Bengio, S. (1996). The vehicle routing problem with time windows – Part II: Genetic search. *INFORMS Journal on Computing* 8, 165–172.

Potvin, J.-Y., Rousseau, J.-M. (1993). A parallel route building algorithm for the vehicle routing and scheduling problem with time windows. *European Journal of Operational Research* 66, 331–340.

Potvin, J.-Y., Rousseau, J.-M. (1995). An exchange heuristic for routing problems with time windows. *Journal of the Operational Research Society* 46, 1433–1446.

Potvin, J.-Y., Kervahut, T., Garcia, B.L., Rousseau, J.-M. (1996). The vehicle routing problem with time windows – Part I: Tabu search. *INFORMS Journal on Computing* 8, 158–164.

Prins, C. (2004). A simple and effective evolutionary algorithm for the vehicle routing problem. *Computers & Operations Research* 31, 1985–2002.

Ralphs, T.K., Kopman, L., Pulleyblank, W.R., Trotter, Jr., L.E. (2003). On the capacitated vehicle routing problem. *Mathematical Programming B* 94, 343–359.

Rechenberg, I. (1973). *Evolutionsstrategie*. Fromman-Holzboog, Stuttgart, Germany.

Rego, C. (1998). A subpath ejection method for the vehicle routing problem. *Management Science* 44, 1447–1459.

Rego, C., Roucairol, C. (1996). A parallel tabu search algorithm using ejection chains for the vehicle routing problem. In: Osman, I.H., Kelly, J.P. (Eds.), *Meta-Heuristics: Theory and Applications*. Kluwer Academic, Boston, pp. 661–675.

Reimann, M., Rubio, R., Wein, L.M. (1999). Heavy traffic analysis of the dynamic stochastic inventory-routing problem. *Transportation Science* 33, 361–380.

Reimann, M., Doerner, K., Hartl, R.F. (2004). D-ants: Savings based ants divide and conquer for the vehicle routing problem. *Computers & Operations Research* 31, 563–591.

Renaud, J., Boctor, F.F., Laporte, G. (1996a). A fast composite heuristic for the symmetric traveling salesman problem. *INFORMS Journal on Computing* 8, 134–143.

Renaud, J., Boctor, F.F., Laporte, G. (1996b). An improved petal heuristic for the vehicle routing problem. *Journal of the Operational Research Society* 47, 329–336.

Rochat, Y., Taillard, É.D. (1995). Probabilistic diversification and intensification in local search for vehicle routing. *Journal of Heuristics* 1, 147–167.

Rossi, F.A., Gavioli, I. (1987). Aspects of heuristic methods in the "Probabilistic traveling salesman problem". In: Andreatta, G., Mason, F., Serafini, P. (Eds.), *Stochastics in Combinatorial Optimization*. World Scientific, Singapore, pp. 214–227.

Russell, R.A. (1977). An effective heuristic for the M-tour traveling salesman problem with some side conditions. *Operations Research* 25, 517–524.

Russell, R.A. (1995). Hybrid heuristics for the vehicle routing problem with time windows. *Transportation Science* 29, 156–166.

Ryan, D.M., Hjorring, C., Glover, F. (1993). Extensions of the petal method for vehicle routing. *Journal of Operational Research Society* 44, 289–296.

Savelsbergh, M.W.P. (1985). Local search in routing problems with time windows. *Annals of Operations Research* 4, 285–305.

Savelsbergh, M.W.P. (1990). En efficient implementation of local search algorithms for constrained routing problems. *European Journal of Operational Research* 47, 75–85.

Savelsbergh, M.W.P. (1992). The vehicle routing problem with time windows: Minimizing route duration. *ORSA Journal on Computing* 4, 146–154.

Savelsbergh, M.W.P., Song, J.-H. (2005). Inventory routing with continuous moves. *Computers & Operations Research*, in press.

Schulze, J., Fahle, T. (1999). A parallel algorithm for the vehicle routing problem with time window constraints. *Annals of Operations Research* 86, 585–607.

Secomandi, N. (1998). Exact and heuristic dynamic programming algorithms for the vehicle routing problem with stochastic demands. PhD dissertation, Faculty of the College of Business Administration, University of Houston, TX.

Secomandi, N. (2000). Comparing neuro-dynamic programming algorithms for the vehicle routing problem with stochastic demands. *Computers & Operations Research* 27, 1201–1225.

Secomandi, N. (2003). Analysis of a rollout approach to sequencing problems with stochastic routing applications. *Journal of Heuristics* 9, 321–352.

Séguin, R. (1994). Problèmes stochastiques de tournées de véhicules. PhD thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal, Canada.

Semet, F., Taillard, É.D. (1993). Solving real-life vehicle routing problems efficiently using tabu search. *Annals of Operations Research* 41, 469–488.

Shaw, P. (1998). Using constraint programming and local search methods to solve vehicle routing problems. In: Maher, M., Puget, J.-F. (Eds.), *Principles and Practice of Constraint Programming*. Springer-Verlag, New York, pp. 417–431.

Sniedovich, M. (1981). Analysis of a preference order traveling salesman problem. *Operations Research* 29, 1234–1237.

Solomon, M.M. (1987). Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research* 35, 254–265.

Solomon, M.M., Baker, E.K., Schaffer, J.R. (1988). Vehicle routing and scheduling problems with time window constraints: Efficient implementations of solution improvement procedures. In: Golden, B.L., Assad, A.A. (Eds.), *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, pp. 85–106.

Song, J.-H., Savelsbergh, M.W.P. (2005). Performance measurement for inventory routing. *Transportation Science*, in press.

Stewart, W.R., Golden, B.L. (1983). Stochastic vehicle routing: A comprehensive approach. *European Journal of Operational Research* 14, 371–385.
Taillard, É.D. (1993). Parallel iterative search methods for vehicle routing problems. *Networks* 23, 661–673.
Taillard, É.D., Badeau, P., Gendreau, M., Guertin, F., Potvin, J.-Y. (1997). A tabu search heuristic for the vehicle routing problem with soft time windows. *Transportation Science* 31, 170–186.
Tan, K.C., Lee, L.H., Ou, K. (2001). Hybrid genetic algorithms in solving vehicle routing problems with time window constraints. *Asia-Pacific Journal of Operational Research* 18, 170–186.
Tarantilis, C.-D., Kiranoudis, C.T. (2002). Bone route: Adaptive memory method for effective fleet management. *Annals of Operations Research* 115, 227–241.
Thangiah, S.R., Petrovic, P. (1998). Introduction to genetic heuristics and vehicle routing problems with complex constraints. In: *Advances in Computational and Stochastic Optimization, Logic Programming, and Heuristic Search*. *Operations Research/Computer Science Interfaces*, vol. 9. Kluwer Academic, Boston, pp. 253–286.
Thompson, P.M., Psaraftis, H.N. (1993). Cyclic transfer algorithms for multi-vehicle routing and scheduling problems. *Operations Research* 41, 935–946.
Tillman, F. (1969). The multiple terminal delivery problem with probabilistic demands. *Transportation Science* 3, 192–204.
Toth, P., Vigo, D. (1995). An exact algorithm for the capacitated shortest spanning arborescence. *Annals of Operations Research* 61, 121–142.
Toth, P., Vigo, D. (1997). An exact algorithm for the vehicle routing problem with backhauls. *Transportation Science* 31, 372–385.
Toth, P., Vigo, D. (1998). Exact algorithms for vehicle routing. In: Crainic, T., Laporte, G. (Eds.), *Fleet Management and Logistics*. Kluwer Academic, Boston, pp. 1–31.
Toth, P., Vigo, D. (Eds.) (2002a). *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia.
Toth, P., Vigo, D. (2002b). An overview of vehicle routing problems. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 1–26.
Toth, P., Vigo, D. (2002c). Branch-and-bound algorithms for the capacitated VRP. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, pp. 29–51.
Toth, P., Vigo, D. (2002d). Models, relaxations and exact approaches for the capacitated vehicle routing problem. *Discrete Applied Mathematics* 123, 487–512.
Toth, P., Vigo, D. (2003). The granular tabu search and its application to the vehicle routing problem. *INFORMS Journal on Computing* 15, 333–346.
Trudeau, P., Dror, M. (1992). Stochastic inventory routing: Route design with stockouts and route failures. *Transportation Science* 26, 171–184.
Van Breedam, A. (1994). An analysis of the behavior of heuristics for the vehicle routing problem for a selection of problems with vehicle-related, customer-related, and time-related constraints. PhD dissertation, University of Antwerp, Belgium.
Van Slyke, R., Wets, R.J.-B. (1969). *L*-shaped programs with applications to optimal control and stochastic programming. *SIAM Journal of Applied Mathematics* 17, 638–653.
Verweij, B., Ahmed, S., Kleywegt, A.J., Nemhauser, G.L., Shapiro, A. (2003). The sample average approximation method applied to stochastic routing problems: A computational study. *Computational Optimization and Applications* 24, 289–333.
Vigo, D. (1996). A heuristic algorithm for the asymmetric capacitated vehicle routing problem. *European Journal of Operational Research* 89, 108–126.
Viswanathan, S., Mathur, K. (1997). Integrating routing and inventory decisions in one-warehouse multiretailer multiproduct distribution systems. *Management Science* 43, 294–312.
Volgenant, A., Jonker, R. (1983). The symmetric traveling salesman problem and edge exchange in minimal 1-trees. *European Journal of Operational Research* 12, 395–403.
Voudouris, C. (1997). Guided local search for combinatorial problems. Dissertation, University of Essex, United Kingdom.

Waller, M., Johnson, M.E., Davis, T. (1999). Vendor-managed inventory in the retail supply chain. *Journal of Business Logistics* 20, 183–203.

Wang, X., Regan, A.C. (2001). Assignment models for local truckload trucking problems with stochastic service times and time window constraints. *Transportation Research Record* 1171, 61–68.

Wark, P., Holt, J. (1994). A repeated matching heuristic for the vehicle routing problem. *Journal of Operational Research Society* 45, 1156–1167.

Waters, C.D.J. (1989). Vehicle-scheduling problems with uncertainty and omitted customers. *Journal of the Operational Research Society* 40, 1099–1108.

Webb, R., Larson, R. (1995). Period and phase of customer replenishment: A new approach to the strategic inventory/routing problem. *European Journal of Operational Research* 85, 132–148.

Willard, J.A.G. (1989). Vehicle routing using *r*-optimal tabu search. MSc dissertation, The Management School, Imperial College, London.

Wren, A. (1971). *Computers in Transport Planning and Operation*. Ian Allan, London.

Wren, A., Holliday, A. (1972). Computer scheduling of vehicles from one or more depots to a number of delivery points. *Operational Research Quarterly* 23, 333–344.

Xu, J., Kelly, J.P. (1996). A network flow-based tabu search heuristic for the vehicle routing problem. *Transportation Science* 30, 379–393.

Yang, W.H., Mathur, K., Ballou, R.H. (2000). Stochastic vehicle routing with restocking. *Transportation Science* 34, 99–112.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14007-4

Chapter 7

Transportation on Demand

Jean-François Cordeau

Canada Research Chair in Logistics and Transportation, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal, H3T 2A7, Canada
E-mail: Jean-Francois.Cordeau@hec.ca

Gilbert Laporte

Canada Research Chair in Distribution Management, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal, H3T 2A7, Canada
E-mail: gilbert@crt.umontreal.ca

Jean-Yves Potvin

Département d'informatique et de recherche opérationnelle and Centre de recherche sur les transports, Université de Montréal C.P. 6128, succ. Centre-Ville, Montréal, H3C 3J7, Canada
E-mail: potvin@iro.umontreal.ca

Martin W.P. Savelsbergh

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA
E-mail: mwps@isye.gatech.edu

1 Introduction

Transportation on Demand (TOD) is concerned with the transportation of passengers or goods between specific origins and destinations at the request of users. Common examples are dial-a-ride transportation services for the elderly and the disabled, urban courier services, aircraft sharing, and emergency vehicle dispatching. In all such systems, users formulate requests for transportation from a pickup point to a delivery (or drop-off) point. These requests are served by a set of capacitated vehicles that often provide a shared service in the sense that several passengers or goods may be in a vehicle at the same time.

In recent years, TOD systems have become increasingly popular for a number of reasons. With the aging of the population and the trend toward the development of ambulatory health care services, more and more people rely on door-to-door transportation systems provided by local authorities. Aircraft sharing has also gained in popularity thanks to cost reduction efforts made by organizations and to the numerous problems that have recently plagued the airline industry. Finally, a growing emphasis on electronic commerce, cycle-time compression, and just-in-time deliveries has increased the need for demand-responsive freight transportation systems.

TOD systems can be either static or dynamic. In the first case, all requests are known beforehand while, in the second case, requests are received dynamically and vehicle routes must be adjusted in real-time to meet demand. For instance, courier services are generally highly dynamic whereas dial-a-ride systems can be regarded as mostly static since they usually require users to make a reservation at least one day in advance. In practice, dynamic problems are often treated as sequences of static subproblems. Reoptimization from the current solution can be performed whenever a new request is formulated, or requests can be buffered and periodically incorporated in the existing vehicle routes in batches.

Most TOD problems are characterized by the presence of three often conflicting objectives: maximizing the number of requests served, minimizing operating costs, and minimizing user inconvenience. A balance between these objectives is sometimes obtained by first maximizing the number of requests that can be accepted given the available capacity and then minimizing the operating costs while imposing service quality constraints. Service quality is usually measured in terms of deviations from desired pickup and delivery times and, in the case of passenger transportation, in terms of excess ride time (i.e., the difference between the actual ride time of a user and the minimum possible ride time). Operating costs are mostly related to the number of vehicles used, to total route duration and to total distance traveled by the vehicles.

Another distinguishing aspect of TOD problems is the importance of the temporal dimension. Pickups and deliveries are often restricted to take place within specified time windows. These time windows are sometimes very narrow, especially in the case of passenger transportation. In this context, quality of service is also often controlled by imposing a limit on the ride time of each user. The latter is particularly important in the case of emergency vehicles. Waiting while passengers are in a vehicle can also be prohibited. Finally, maximum route duration constraints are sometimes imposed to take driver shift lengths into account.

The day-to-day management of a TOD system involves making decisions regarding three main aspects: request clustering, vehicle routing, and vehicle scheduling. Request clustering consists of creating groups of requests to be served by the same vehicle because of their spatial and temporal proximity. Given these groups, vehicle routing consists of deciding the order in which the associated pickup and delivery locations should be visited by each vehicle. Finally, vehicle scheduling specifies the exact time at which each location should be visited. These decisions are obviously tightly intertwined and a proper management of the system calls for their simultaneous optimization.

The Operations Research literature contains numerous studies addressing both static and dynamic TOD problems. Most variants are, in fact, generalizations of the Vehicle Routing Problem with Pickup and Delivery (VRPPD). The aim of this chapter is to present the most important results regarding the VRPPD and to survey four areas of applications: the dial-a-ride problem, the urban courier service problem, the dial-a-flight problem, and the emergency

vehicle dispatch problem. In the latter application, the main emphasis is on locational issues as opposed to routing choices.

The remainder of the chapter is organized as follows. The next section formally defines the VRPPD, introduces notation that will be used throughout the chapter, and reviews the related literature. The following four sections then each focus on a specific application by describing the particularities of the problem and summarizing the main exact and heuristic solution algorithms that have been proposed in the literature.

2 The vehicle routing problem with pickup and delivery

The VRPPD is a generalization of the classical VRP which also belongs to a larger family of Pickup and Delivery Problems (PDPs). One can distinguish between three well-known types of pickup and delivery problems that have been studied in the literature. One is the single-commodity PDP in which a single type of goods is either picked up or delivered at each node (see, e.g., Hernández-Pérez and Salazar-González, 2004). This is the case, for example, when an armored vehicle transports money between the branch offices of a bank. Another variant is the two-commodity PDP where two types of goods are considered and each node may act as both a pickup and a delivery node (see, e.g., Gendreau et al., 1999; Baldacci et al., 2003). This problem arises, for instance, in beer or soft drinks delivery where vehicles deliver full bottles and collect empty ones. A variant of this problem is the VRP with backhauls in which all deliveries must be performed before any pickup. Finally, the n-commodity problem occurs when each commodity is associated with a single pickup node and a single delivery node. This is the case when passengers or goods must be transported from an origin to a destination. This problem is usually referred to as the VRPPD.

Because most practical applications of the VRPPD include restrictions on the time at which each location may be visited by a vehicle, it is convenient to present a slightly more general variant of the problem, called the VRPPD with Time Windows (VRPPDTW).

Let n denote the number of requests to be satisfied. Assuming that all vehicles are based at a single depot, the VRPPDTW may be defined on a directed graph $G = (N, A)$ where $N = P \cup D \cup \{0, 2n + 1\}$, $P = \{1, \ldots, n\}$ and $D = \{n + 1, \ldots, 2n\}$ are node sets, and $A = \{(i, j)\colon i, j \in N\}$ is the arc set. Subsets P and D contain pickup and delivery nodes, respectively, while nodes 0 and $2n + 1$ represent the origin and destination depots. With each request i are thus associated an origin node i and a destination node $n + i$. Let K be the set of vehicles and let $m = |K|$. Each vehicle $k \in K$ has a capacity Q_k and the total duration of its route cannot exceed T_k. With each node $i \in N$ are associated a load q_i and a nonnegative service duration d_i such that $q_0 = q_{2n+1} = 0$, $q_i = -q_{n+i}$ $(i = 1, \ldots, n)$, and $d_0 = d_{2n+1} = 0$. A time window $[e_i, l_i]$ is also associated with each node $i \in N$, where e_i and l_i represent the earliest and

latest time, respectively, at which service may begin at node i. With each arc $(i, j) \in A$ are associated a routing cost c_{ij} and a travel time t_{ij}.

For each arc $(i, j) \in A$ and each vehicle $k \in K$, let $x_{ij}^k = 1$ if and only if vehicle k travels from node i to node j. For each node $i \in N$ and each vehicle $k \in K$, let B_i^k be the time at which vehicle k begins service at node i, and Q_i^k be the load of vehicle k after visiting node i. The VRPPDTW can be formulated as the following mixed-integer program:

$$\text{minimize} \quad \sum_{k\in K}\sum_{i\in N}\sum_{j\in N} c_{ij}^k x_{ij}^k \tag{1}$$

subject to

$$\sum_{k\in K}\sum_{j\in N} x_{ij}^k = 1, \quad i \in P, \tag{2}$$

$$\sum_{j\in N} x_{ij}^k - \sum_{j\in N} x_{n+i,j}^k = 0, \quad i \in P, k \in K, \tag{3}$$

$$\sum_{j\in N} x_{0j}^k = 1, \quad k \in K, \tag{4}$$

$$\sum_{j\in N} x_{ji}^k - \sum_{j\in N} x_{ij}^k = 0, \quad i \in P \cup D, k \in K, \tag{5}$$

$$\sum_{i\in N} x_{i,2n+1}^k = 1, \quad k \in K, \tag{6}$$

$$B_j^k \geqslant \left(B_i^k + d_i + t_{ij}\right)x_{ij}^k, \quad i \in N, j \in N, k \in K, \tag{7}$$

$$Q_j^k \geqslant \left(Q_i^k + q_j\right)x_{ij}^k, \quad i \in N, j \in N, k \in K, \tag{8}$$

$$B_i^k + d_i + t_{i,n+i} \leqslant B_{n+i}^k, \quad i \in P, k \in K, \tag{9}$$

$$B_{2n+1}^k - B_0^k \leqslant T_k, \quad k \in K, \tag{10}$$

$$e_i \leqslant B_i^k \leqslant l_i, \quad i \in N, k \in K, \tag{11}$$

$$\max\{0, q_i\} \leqslant Q_i^k \leqslant \min\{Q_k, Q_k + q_i\}, \quad i \in N, k \in K, \tag{12}$$

$$x_{ij}^k \in \{0, 1\}, \quad i \in N, j \in N, k \in K. \tag{13}$$

The objective function minimizes the total routing cost. Constraints (2) and (3) ensure that each request is served exactly once and that the associated pickup and delivery nodes are visited by the same vehicle. Constraints (4)–(6) guarantee that the route of each vehicle k starts at the origin depot and ends at the destination depot. The consistence of time and load variables is ensured by constraints (7) and (8). Constraints (9) force the vehicles to visit the pickup node of a request before its delivery node. Finally, inequalities (10) bound the duration of each route while (11) and (12) impose time windows and capacity constraints, respectively.

The VRPPDTW is NP-hard since it generalizes the Traveling Salesman Problem (TSP) known to be NP-hard (Garey and Johnson, 1979). In the presence of time windows, even finding a feasible solution to the problem is NP-hard since the feasibility problem for the TSP with time windows is itself NP-complete (Savelsbergh, 1985).

Savelsbergh and Sol (1995) considered a slightly more general formulation of the pickup and delivery problem and reviewed the relevant literature on the problem. A more recent survey on pickup and delivery problems was also prepared by Desaulniers et al. (2002). In the remainder of this section, we review the most important exact and heuristic solution algorithms for the VRPPD with and without time windows. We first present algorithms for the single-vehicle case, followed by the multiple-vehicle case.

2.1 The single-vehicle VRPPD

The single-vehicle VRPPD is obtained when $m = 1$ in formulation (1)–(13). Although few real-life applications exist for this problem, it may appear as a subproblem in algorithms for the multiple-vehicle case. It is worth mentioning that unlike the single-vehicle VRP which necessarily reduces to an (uncapacitated) TSP, the single-vehicle VRPPD may incorporate capacity constraints. Indeed, because both pickup and delivery nodes are considered, any number of requests may be served by a single vehicle provided that $q_i \leqslant Q$ for every request.

2.1.1 Exact algorithms

Kalantari et al. (1985) have presented branch-and-bound algorithms for the single-vehicle case with finite and infinite vehicle capacity. These algorithms, which are modifications of the algorithm of Little et al. (1963) for the TSP, work by eliminating in each branch of the search tree all arcs that would lead to a violation of a precedence constraint. Fischetti and Toth (1989) have developed an additive bounding procedure for the more general TSP with precedence constraints in which some nodes may have one or several predecessors. This procedure combines the lower bounds obtained from the assignment problem and shortest spanning 1-arborescence problem relaxations, variable decomposition and disjunctions. A lower bounding procedure and a dynamic programming algorithm were later developed by Bianco et al. (1994) while Balas et al. (1995) proposed valid inequalities for this problem.

More recently, Ruland and Rodin (1997) introduced a branch-and-cut algorithm for the TSP with Pickup and Delivery (TSPPD). Using the previously introduced notation, the problem can be formulated on an undirected graph $G' = (N, E)$ with binary edge variables x_e, $e \in E$.

Given a node set $S \subseteq N$, denote by $E(S)$ and $\delta(S)$, the sets of edges with both endpoints in S, and with one endpoint in S and the other in $N \setminus S$, respectively. Let also $x(S) = \sum_{e \in E(S)} x_e$. Similarly, let $x(E') = \sum_{e \in E'} x_e$ for any edge set $E' \subseteq E$. Finally, let $\mathcal{U}_s = \{U \subset N \mid 0 \in U\}$ and $\mathcal{U}_p = \{U \subset N \mid$

$0 \in U, 2n+1 \notin U, \exists i \in P, i \notin U, n+i \in U\}$. The formulation can be stated as follows:

$$\text{minimize} \quad \sum_{e \in E} c_e x_e \tag{14}$$

subject to

$$x\big(\{0, 2n+1\}\big) = 1, \tag{15}$$

$$x\big(\delta(\{i\})\big) = 2, \quad i \in N, \tag{16}$$

$$x\big(\delta(U)\big) \geqslant 2, \quad U \in \mathcal{U}_{\mathrm{s}}, \tag{17}$$

$$x\big(\delta(U)\big) \geqslant 4, \quad U \in \mathcal{U}_{\mathrm{p}}, \tag{18}$$

$$0 \leqslant x_e \leqslant 1, \quad e \in E, \tag{19}$$

$$x \in \mathbb{Z}^{|E|}. \tag{20}$$

Constraint (15) simply connects the origin depot to the destination depot and ensures that the solution is a Hamiltonian cycle. Each node is then required to have a degree of 2 by constraints (16) while constraints (17) ensure the biconnectedness of the solution. Finally, constraints (18) force the pickup node of each request to be visited before its delivery node.

As explained by Ruland (1995) a careful analysis of constraints (17) and (18) reveals that the cardinality of the sets $\mathcal{U}_{\mathrm{s}}$ and $\mathcal{U}_{\mathrm{p}}$ can, in fact, be reduced by exploiting the redundancy of some of the associated constraints. The author also shows that the resulting subtour elimination constraints and precedence constraints define faces of the TSPPD polytope.

Two other classes of inequalities were introduced by Ruland (1995). Let $U_1, \ldots, U_m \subset N$ be mutually disjoint subsets and let $i_1, \ldots, i_m \in P$ be requests such that $0, 2n+1 \notin U_l$ and $i_l, n+i_{l+1} \in U_l$ for $l = 1, \ldots, m$ (where $i_{m+1} = i_1$). The following inequality, called a *generalized order constraint*, defines a proper face of the TSPPD polytope:

$$\sum_{l=1}^{m} x(U_l) \leqslant \sum_{l=1}^{m} |U_l| - m - 1. \tag{21}$$

(Note that similar inequalities were also proposed by Balas et al. (1995) for the precedence-constrained asymmetric TSP.)

Consider two nodes $i, j \in P$ and a subset H such that $\{i, j\} \subseteq H \subseteq N \setminus \{0, n+i, n+j, 2n+1\}$. The following inequality, called an *order matching constraint*, also defines a proper face of the TSPPD polytope:

$$x(H) + x\big(\{i, n+i\}\big) + x\big(\{j, n+j\}\big) \leqslant |H|. \tag{22}$$

The latter inequality can, in fact, be lifted by considering all requests p for which $p \in H$ and $n+p \in N \setminus H$. For each type of inequality, Ruland (1995) described separation algorithms relying on the solution of maximum flow problems. Computational results were reported on instances with $n \leqslant 15$.

For the single-vehicle VRPPD with time windows and capacity constraints, Desrosiers et al. (1986) have developed an exact forward dynamic programming algorithm to minimize the total distance traveled. A state (S, i) is defined if there exists a feasible path that starts at the depot 0, visits all nodes in $S \subseteq N$ and ends at node $i \in S$. For each such state, two-dimensional labels are used to keep track of the time and distance traveled. A label can be eliminated if there exists no feasible path that starts at node i and visits all remaining nodes. Computational experiments performed on real-life data with tight time windows have shown that the algorithm could very quickly solve instances with $n \leqslant 40$.

2.1.2 Heuristics

A probabilistic analysis of a simple construction heuristic for the problem without capacity and time windows was performed by Stein (1978). This heuristic constructs a solution by concatenating two optimal traveling salesman tours: one through the n origins and one through the n destinations. The author showed that if the $2n$ points are drawn independently from the uniform probability distribution over a subset of the Euclidean plane, then the algorithm has an asymptotic performance bound of 1.06. Later, Psaraftis (1983a) presented a worst-case analysis of a two-phase construction heuristic. In the first phase, an optimal TSP tour is constructed for the $2n$ points. In the second phase, a solution to the pickup and delivery problem is obtained by traversing the TSP tour clockwise until all points are visited. While doing this, points that have already been visited or that correspond to a destination whose origin has not been visited should be skipped. Psaraftis showed that if the minimum spanning tree heuristic of Christofides (1976) is used to construct the TSP tour, the heuristic has a worst-case performance ratio of 3.0. He also reported computational experiments indicating that on realistic size instances the average performance of his heuristic was superior to that of Stein's heuristic. In a related paper, Psaraftis (1983c) proposed a local search heuristic that extends the TSP interchange procedure of Lin (1965) to handle precedence constraints. In addition, Psaraftis described an approach that identifies the best k-interchange in $O(n^k)$ time. Similar ideas were introduced by Savelsbergh (1990) in the more general context of constrained routing problems. Healy and Moll (1995) also have described a variant of local search for the same problem. Their strategy, called *sacrificing*, consists of biasing the search in the direction of solutions with larger neighborhoods of feasible solutions in the hope of improving the overall quality of the local optima found.

Renaud et al. (2000) have later described adaptations of some classical TSP heuristics to handle the pickup and delivery problem. They have also introduced a two-phase method in which the first phase constructs a solution by means of a double insertion procedure that performs the simultaneous insertion of a pickup node and its associated delivery node, and the second phase is a deletion and reinsertion improvement procedure based on 4-opt* heuristic of Renaud et al. (1996). Recently, Renaud et al. (2002) have described and compared several perturbation heuristics whose aim is to help an improve-

ment algorithm to escape from a local optimum. In particular, they considered instance perturbation in which the instance data are slightly changed, algorithmic perturbation in which the definition of the neighborhood is modified, and solution perturbation in which a locally optimal solution is perturbed before restarting the search. Compared with existing heuristics, these perturbation schemes yielded excellent results on instances with $n \leqslant 220$.

For the single-vehicle problem with time windows, Van der Bruggen et al. (1993) developed a two-phase local search method based on the variable-depth search of Lin and Kernighan (1973) for the TSP. In the first phase, nodes are first sorted in increasing order of the middle of their time window. The resulting ordering is then modified so as to ensure that the pickup node of each request appears before the delivery node and capacity constraints are satisfied. A solution is then constructed by visiting the nodes in that order. This solution may violate some of the time windows. Iterative improvements are then performed in the hope of obtaining a feasible solution. This solution is also further refined by applying the same exchange procedures with a different objective.

2.2 The multiple-vehicle VRPPD

2.2.1 Exact algorithms

Dumas et al. (1991) have proposed a set-partitioning formulation of the problem and an exact column generation algorithm. For vehicle $k \in K$, let Ω^k be the set of feasible routes and let c_r^k be the cost of route r. In addition to traditional flow conservation constraints, each route $r \in \Omega^k$ satisfies time windows, capacity constraints, pairing constraints (i.e., node $i \in P$ is visited iff node $n+i \in D$ is also visited by the route), and precedence constraints. For all $i \in P$ and $r \in \Omega^k$, let a_{ir}^k be a binary constant equal to 1 if request i is served by route r of vehicle k, and 0 otherwise. Finally, define a binary variable y_r^k that takes the value 1 if route r is used for vehicle k, and 0 otherwise. The problem can be stated as follows:

$$\text{minimize} \quad \sum_{k \in K} \sum_{r \in \Omega^k} c_r^k y_r^k \tag{23}$$

subject to

$$\sum_{k \in K} \sum_{r \in \Omega^k} a_{ir}^k y_r^k = 1, \quad \forall i \in P, \tag{24}$$

$$\sum_{r \in \Omega^k} y_r^k = 1, \quad \forall k \in K, \tag{25}$$

$$y_r^k \in \{0, 1\}, \quad \forall k \in K, r \in \Omega^k. \tag{26}$$

This formulation is solved by a branch-and-bound method in which the linear relaxations are solved by column generation. Columns of negative reduced cost are generated by solving a resource-constrained shortest path problem

in which the arc costs are modified to reflect the current values of the dual variables associated with constraints (24) and (25). This problem is solved by a dynamic programming algorithm which is very similar to the one described by Desrosiers et al. (1986) for the single-vehicle case. In this case, however, not all nodes have to be visited by the vehicle. To obtain integer solutions, branching is performed on additional order variables O_{ij}, $i, j \in P \cup \{0, 2n+1\}$, indicating the sequence in which pickups are performed. These decisions are easily transfered to the subproblem and are handled directly by the dynamic programming algorithm through the introduction of an additional label representing the last pickup node visited. Several arc elimination rules are proposed by the authors to reduce the problem size by taking time windows and pairing constraints into account. For example, arc $(i, n+j)$ can be eliminated if the path $j \to i \to n+j \to n+i$ is infeasible even when setting $B_j = e_j$. The algorithm was successful in solving two real-life instances with 19 and 30 requests, respectively, as well as randomly generated instances involving up to 55 requests but tight capacity constraints. According to the authors, the algorithm works well when capacity constraints are restrictive and each route serves a small number of requests (i.e., five or fewer).

A similar approach was developed by Savelsbergh and Sol (1998). However, because it was intended to solve large-scale instances, it differs from that of Dumas et al. (1991) in the following respects:

(i) whenever possible, construction and improvement heuristics are used to solve the pricing subproblem;

(ii) a sophisticated column management mechanism scheme is used to keep the column generation master problem as small as possible;

(iii) columns are selected with a bias toward increasing the likelihood of identifying feasible integer solutions during the solution of the master problem;

(iv) branching decisions are made on additional assignment variables z_i^k representing the fraction of request i that is served by vehicle k; and

(v) a primal heuristic is used at each node of the search tree to obtain upper bounds.

Computational results performed by the authors on instances with $n \leqslant 50$ show that the proposed approach yields high quality solutions in short computing times even when capacity constraints are not very tight.

Very recently, another column generation method was used by Xu et al. (2003) to address a complex pickup and delivery problem encountered in long-haul transportation planning. In their problem, there are multiple carriers and multiple-vehicle types available to cover a set of pickup and delivery requests, each of which has multiple pickup time windows and multiple delivery time windows. In addition to the classical vehicle capacity, route duration, pairing and precedence constraints, vehicle routes must satisfy compatibility constraints between the requests, the carriers, and the vehicle types, as well as sequencing constraints requiring the goods to be collected and delivered in a

last-in first-out sequence, i.e., the goods picked-up last must be the first to be delivered. Constraints regarding maximum driving time and maximum working time are also taken into account, leading to a complex objective function incorporating fixed costs, mileage costs, waiting costs, and layover (driver rest) costs. This problem is solved by means of a column generation approach in which the pricing subproblems are solved by fast heuristics. Instead of embedding the column generation method in a branch-and-bound process, the method reaches an integer solution by applying an IP solver to the restricted set of columns generated in solving the linear relaxation of the problem. Comparisons with lower bounds obtained by solving the LP relaxation of the problem exactly through the use of dynamic programming for the pricing subproblem show that the heuristic approaches are capable of generating near-optimal solutions quickly for randomly generated instances with up to 200 requests. Results are also reported on larger instances involving 500 requests.

2.2.2 Heuristics

A tabu search heuristic for the pickup and delivery problem with time windows was developed by Nanry and Barnes (2000). Solutions that violate time window and vehicle capacity constraints are allowed during the search. These authors have considered three types of move. The first removes a node pair $(i, n + i)$ from its current route and reinserts it in a different route. The second swaps two pairs of nodes between two distinct routes. The last consists of moving a single node within its current route. A hierarchical search mechanism is used to dynamically alternate between these neighborhoods according to problem difficulty. Computational results are reported on random instances involving up to 100 requests. A similar tabu search heuristic was also developed by Lau and Liang (2002).

More recently, Ropke and Pisinger (2004) introduced an adaptive Large Neighborhood Search (LNS) heuristic for the VRPPD with time windows. Instead of relying on operators that perform minor changes to the solution in every iteration, the LNS heuristic uses large moves that can potentially rearrange up to 30–40% of all requests in a single iteration. This is accomplished by using several large neighborhoods in an adaptive way. Three removal heuristics are considered: removing random requests, removing similar requests likely to be interchangeable in the solution, and removing requests whose removal yields a large decrease in solution cost. To reinsert these requests, two insertion heuristics are used: a greedy heuristic that inserts at each iteration the request with the least insertion cost, and a regret heuristic that takes into account the cost of not being able to insert a request in its least-cost route. The selection of removal and insertion heuristics is performed randomly at each iteration according to a probability distribution whose weights are adjusted dynamically throughout the search. In experiments performed on test instances with $n \leqslant 500$, the adaptive LNS algorithm outperformed previously proposed heuristics.

3 The dial-a-ride problem

The Dial-a-Ride Problem (DARP) is a generalization of the VRPPD arising in contexts where passengers are transported, either in groups or individually, between specified origins and destinations. The most common DARP application arises in door-to-door transportation services for elderly or handicapped people. In this context, users often formulate two requests per day: an *outbound* request from home to a destination, and an *inbound* request for the return trip. The DARP distinguishes itself from the basic VRPPD by its focus on controlling user inconvenience. This usually takes the form of constraints or objective function terms relating to waiting time, ride time (i.e., the time spent by a user in the vehicle) as well as deviations from desired departure and arrival times.

In the remainder of this section, we first discuss the scheduling aspect of the problem and then present the most important exact and heuristic algorithms for the static single-vehicle and multiple-vehicle cases, respectively. This is followed by the dynamic case in the last section. An overview of some of these methods can also be found in the Cordeau and Laporte (2003a) survey.

3.1 Scheduling

Because of the focus on controlling user inconvenience, the scheduling aspect of the problem plays a central role in most applications. As a result, the problem of finding an optimal schedule for a given vehicle route has been studied independently in the literature. In general terms, given a sequence of nodes $i_1, i_2, \ldots, i_q$ to be visited, the problem of finding an optimal schedule satisfying time windows can be formulated as

$$\text{minimize} \quad \sum_{i=1}^{q} g_i(B_i) \tag{27}$$

subject to

$$B_i + d_i + t_{i,i+1} \leqslant B_{i+1}, \quad \forall i = 1, \ldots, q-1, \tag{28}$$

$$e_i \leqslant B_i \leqslant l_i, \quad \forall i = 1, \ldots, q, \tag{29}$$

where $g_i(B_i)$ is a convex function defined with respect to the time window $[e_i, l_i]$. Sexton and Bodin (1985a, 1985b) observed that some special cases of this scheduling problem can be seen as network flow problems and thus solved very efficiently. Dumas et al. (1989) proposed a dual approach to solve the general problem by performing q unidimensional minimizations. In the special cases where the inconvenience functions are quadratic or linear, the complexity of the algorithm is $O(q)$. The extension of this methodology to handle time-varying, stochastic travel times was addressed by Fu (2002).

On a related topic, Hunsaker and Savelsbergh (2002) have devised a procedure for efficiently testing the feasibility of an insertion in construction or

improvement heuristics. They considered a variant of the DARP with time windows, waiting time constraints and ride time constraints, and showed how to check in $O(q)$ time whether the insertion of a given request in a route is feasible. Cordeau and Laporte (2003b) also proposed a procedure, based on the *forward time slack* notion introduced by Savelsbergh (1992), to sequentially minimize time window constraint violations, route durations, and ride times in the context of local search.

3.2 The static single-vehicle DARP

Early work on the single-vehicle dial-a-ride problem was carried out by Psaraftis (1980) who studied the "immediate-request" case in which a list of requests should be served as soon as possible. His model assumes that no time windows are specified by the users. Instead the transporter imposes "maximum position shift" constraints limiting the difference between the position of a request in the calling list and its position in the vehicle route. The objective function aims to minimize the sum of route completion time and customer dissatisfaction. Customer dissatisfaction is itself expressed as a weighted combination of waiting time before pickup and ride time. The problem is solved using a dynamic programming algorithm in which the state space consists of vectors $(L, k_1, \ldots, k_n)$, where L denotes the node being currently visited and k_i denotes the status of request i. The status of request i is either 3 if user i has been dropped off, 2 if the user is still in the vehicle or 1 if the user has yet to be picked up. The complexity of this algorithm is $O(n^2 3^n)$ and only small instances can thus be solved. Psaraftis also explains how to handle the dynamic case in which new requests occur dynamically in time but no information on future requests is available. In this context, maximum position shift constraints become essential to prevent a request from being indefinitely deferred. In a later paper, Psaraftis (1983b) extended his approach to handle time windows on departure and arrival times. The new algorithm has the same complexity as the previous one but uses forward instead of backward recursion.

A heuristic approach based on Benders decomposition was later developed by Sexton and Bodin (1985a, 1985b) who considered one-sided time windows on delivery. Their algorithm iterates between a routing master problem and a scheduling subproblem. The routing problem relaxes the Benders cuts in the objective function and is solved by a route improvement procedure. The scheduling subproblem is shown to be the dual of a network flow problem that can be solved very quickly. These authors minimize a user inconvenience function made up of the weighted sum of two terms. The first measures the difference between the actual travel time and the direct travel time of a user. The second term is the (positive) difference between desired drop-off time and actual drop-off time, under the assumption that the former is at least as large as the latter, late drop-offs being disallowed. The approach was tested on real-life data sets where the number of users varies between 7 and 20.

3.3 The static multiple-vehicle DARP

One of the first heuristics for the multiple-vehicle DARP was proposed by Jaw et al. (1986) who impose windows on the pickup times of inbound requests and on the drop-off times of outbound requests. A maximum ride time, expressed as a linear function of the direct ride time, is given for each user. In addition, vehicles are not allowed to be idle when carrying passengers. A nonlinear objective function combining several types of disutility is used to assess the quality of solutions. The authors have developed an insertion heuristic that selects users in order of earliest feasible pickup time and gradually inserts them into vehicle routes so as to yield the least possible increase in the objective function. The algorithm was tested on artificial instances involving 250 users and on a real data set with 2617 users and 28 vehicles. This approach was also adapted by Alfa (1986) and applied to a practical case in Winnipeg, Canada.

Several of the heuristics proposed for the multiple-vehicle case are two-phase algorithms in which the first phase creates and selects clusters of users that are then combined into vehicle routes in the second phase. An early approach based on this idea is the interactive optimizer described by Cullen et al. (1981) for the case of a homogeneous fleet. Another clustering method was proposed by Bodin and Sexton (1986). Their heuristic creates clusters, applies the single-vehicle algorithm of Sexton and Bodin (1985a, 1985b) to each cluster and then moves users between clusters so as to reduce total user inconvenience. It was applied to real-life instances involving approximately 85 users each.

Dumas et al. (1989) later improved upon this methodology by creating so-called "mini-clusters" of users, i.e., groups of users to be served within the same area at approximately the same time. Users in a mini-cluster should be transportable by a single vehicle while respecting constraints on time windows, vehicle capacity, pairing, and precedence. The mini-clusters are then optimally combined to form feasible vehicle routes, using column generation. In this phase, a time window is imposed on each cluster to ensure feasibility and columns are generated by solving a constrained shortest-path problem. Finally, each vehicle route is optimized by means of the single-vehicle algorithm of Desrosiers et al. (1986) and a scheduling step is executed to minimize user inconvenience (Dumas et al., 1990). Instances with up to 200 users are easily solved, while larger instances require the use of a spatial and temporal decomposition technique. The mini-clustering phase was later improved by Desrosiers et al. (1991) who described a parallel insertion method that relies on the notion of neighboring requests. Two requests are said to be neighbors if they satisfy the following conditions:

1. $e_i \leqslant e_j \leqslant l_{n+i}$ or $e_i \leqslant l_{j+n} \leqslant l_{n+i}$ or $e_j \leqslant e_i \leqslant l_{n+i} \leqslant l_{n+j}$;
2. $t_{ij} + t_{j,n+i} \leqslant \alpha t_{i,n+i}$ or $t_{ji} + t_{i,n+j} \leqslant \alpha t_{j,n+j}$ with $\alpha > 1$;
3. $|\theta_i - \theta_j| \leqslant \beta$ where θ_i is the angle between a reference axis and the direction from i to $n + i$;

4. $s(i, j) \geqslant \gamma$ where $s(i, j)$ are the savings in distance obtained by clustering requests i and j together.

Clusters are then constructed in parallel by treating the requests in decreasing order of the direct duration $t_{i,n+i}$, and considering, at each iteration, the creation of a new cluster or the insertion of a request in clusters that contain at least one neighboring request. Finally, Ioachim et al. (1995) showed that there is an advantage in terms of solution quality to use an optimization technique for the construction of the clusters. Results were reported on data sets comprising more than 2500 users.

Another study, by Borndörfer et al. (1999), also uses a two-phase approach in which clusters of users are first constructed and then grouped together to form feasible vehicle routes. A cluster is defined as a "maximal subtour such that the vehicle is never empty". In the first phase, a large set of good clusters are constructed and a set partitioning problem is then solved to select a subset of clusters serving each user exactly once. In the second phase, feasible routes are enumerated by combining clusters and a second set partitioning problem is solved to select the best set of routes covering each cluster exactly once. Both set partitioning problems are solved by a branch-and-cut algorithm. On real-life instances, the algorithm cannot always be run to completion and terminates with the best known solution. It was applied to several instances provided by an operator in Berlin and including between 859 and 1771 transportation requests per day.

In another real-life application, Toth and Vigo (1996) considered a problem in which users specify requests with a time window on their origin or destination. An upper bound proportional to direct distance is imposed on the ride time. Transportation is supplied by a fleet of capacitated minibuses and by the occasional use of taxis. The objective is to minimize the total cost of service. The authors have developed a heuristic consisting of first assigning requests to routes by means of a parallel insertion procedure, and then performing intra-route and inter-route exchanges. Tests performed on instances involving between 276 and 312 requests show significant improvements with respect to the previous hand-made solutions. Further improvements were later obtained by Toth and Vigo (1997) through the execution of a tabu thresholding post-optimization phase after the parallel insertion step.

More recently, Cordeau and Laporte (2003b) have developed a tabu search heuristic for the problem in which users specify a desired arrival time for their outbound trip and a desired departure time for their inbound trip, and a maximum ride time is associated with each user. Capacity and maximum route duration constraints are also imposed on the vehicles. The search algorithm is based on a simple mechanism that iteratively removes a request from its current route and reinserts it into another route. As is common in such contexts (see, e.g., Cordeau et al., 1997), intermediate infeasible solutions are allowed during the search, but they are discouraged by a penalty term in the objective function. As explained in the previous section, whenever the cost of an exchange is evaluated, the schedules of the two routes involved in the exchange

must be updated so as to measure the impacts on violations of time windows, route duration constraints and ride time constraints. The algorithm was tested on randomly generated instances with up to 144 users and on six data sets with $n = 200$ or $n = 295$ provided by a Danish transporter. Parallel implementations of this heuristic have also been described by Attanasio et al. (2004) who studied the case where a fraction of the requests are received dynamically.

Finally, Cordeau (2006) developed a branch-and-cut algorithm for the same version of the problem. This algorithm relies on several types of valid inequalities that are either new inequalities for the problem or adaptations of known inequalities for the TSP, the TSPPD, or the VRP. The first four types of inequalities are liftings of subtour elimination constraints for the symmetric and asymmetric TSP. Consider the simple subtour elimination constraint $x(S) \leqslant |S| - 1$ for $S \subseteq P \cup D$. In the case of the DARP, this inequality can be lifted in two different ways by taking into account the fact that for each user i, node i must be visited before node $n + i$. For any set $S \subseteq P \cup D$, let $\pi(S) = \{i \in P \mid n + i \in S\}$ and $\sigma(S) = \{n + i \in D \mid i \in S\}$ denote the sets of *predecessors* and *successors* of S, respectively. Balas et al. (1995) have proposed two families of inequalities for the precedence-constrained asymmetric TSP that also apply to the DARP by observing that each node $i \in P \cup D$ is either the predecessor or the successor of exactly one other node. For $S \subseteq P \cup D$, the following inequality, called a *successor inequality* (or σ-inequality) is valid for the DARP:

$$x(S) + \sum_{i \in \overline{S} \cap \sigma(S)} \sum_{j \in S} x_{ij} + \sum_{i \in \overline{S} \setminus \sigma(S)} \sum_{j \in S \cap \sigma(S)} x_{ij} \leqslant |S| - 1. \tag{30}$$

Similarly, for any set $S \subseteq P \cup D$, the following *predecessor inequality* (or π-inequality) is valid for the DARP:

$$x(S) + \sum_{i \in S} \sum_{j \in \overline{S} \cap \pi(S)} x_{ij} + \sum_{i \in S \cap \pi(S)} \sum_{j \in \overline{S} \setminus \pi(S)} x_{ij} \leqslant |S| - 1. \tag{31}$$

Directed subtour elimination constraints proposed by Grötschel and Padberg (1985) for the asymmetric TSP can also be lifted in a similar fashion, yielding two other valid inequalities.

Generalized order constraints, introduced by Ruland (1995) for the TSPPD (see inequalities (21)), can also be lifted in two ways as follows:

$$\sum_{l=1}^{m} x(U_l) + \sum_{l=2}^{m-1} x_{i_1, i_l} + \sum_{l=3}^{m} x_{i_1, n+i_l} \leqslant \sum_{l=1}^{m} |U_l| - m - 1, \tag{32}$$

$$\sum_{l=1}^{m} x(U_l) + \sum_{l=2}^{m-2} x_{n+i_1, i_l} + \sum_{l=2}^{m-1} x_{n+i_1, n+i_l} \leqslant \sum_{l=1}^{m} |U_l| - m - 1. \tag{33}$$

Finally, ride time constraints may give rise to paths that are infeasible in an integer solution but nonetheless feasible in a fractional solution. Forbidding such paths can be accomplished as follows. For any directed path $\mathcal{P} =$

$\{i, k_1, k_2, \ldots, k_p, n+i\}$ such that $t_{i,k_1} + d_{k_1} + t_{k_1,k_2} + d_{k_2} + \cdots + t_{k_p,n+i} > L$ the following inequality is valid for the DARP:

$$x_{i,k_1} + \sum_{h=1}^{p-1} x_{k_h,k_{h+1}} + x_{k_p,n+i} \leqslant p - 1. \tag{34}$$

Heuristic separation algorithms are proposed for each type of valid inequality. In addition, several techniques are described to strengthen the formulation and reduce problem size. In computational experiments, the branch-and-cut algorithm was able to solve instances with up to four vehicles and 32 users.

3.4 The dynamic DARP

While early studies on the DARP were often motivated by dynamic settings, the dynamic DARP has received less attention in the literature than its static counterpart. An approach inspired by the work of Jaw et al. (1986) was developed by Madsen et al. (1995) for a real-life problem involving services to elderly and disabled people in Copenhagen. Users may specify a desired pickup or drop-off time window, but not both. Vehicles of several types are used to provide service, not all of which are available at all times. In addition, some requests arrive dynamically throughout the day. New requests are inserted in vehicle routes taking into account their difficulty of insertion into an existing route. The algorithm was tested on a 300-customer, 23-vehicle instance and the authors report that it was capable of generating good quality solutions within very short computing times.

At about the same time, Dial (1995) introduced the concept of an Autonomous Dial-a-Ride Transit (ADART) service based on fully automated command-and-control, order-entry, and routing and scheduling systems implemented on computers on-board vehicles. Here, the system is fully automated: the only human intervention in the process is the customer requesting service. Furthermore, routing-and-scheduling is not done at some central dispatching center, but is rather distributed among vehicles through an auction mechanism.

Teodorovic and Radivojevic (2000) have later studied a generic version of the dynamic dial-a-ride problem using fuzzy logic. Their approach exploits the fact that passengers, dispatchers and drivers have a fuzzy notion of travel times, which can thus be expressed with fuzzy sets and numbers. Through fuzzy arithmetic, calculations about arrival times at customers, waiting times, etc. are performed and the qualitative results are provided to different approximate reasoning algorithms to decide about the assignment and insertion of a new request in a vehicle route.

A software system for demand-responsive passenger services, like variably routed buses, conventional and maxi-taxis, was proposed by Horn (2002). The optimization capabilities of the system are based on least-cost insertions of new requests and periodic reoptimization of the planned routes. The latter is a steepest descent approach using a neighborhood structure which either moves

or swaps customers. A so-called "rank-homing" heuristic is also proposed for governing the relocation of idle vehicles. A set of locations, known as cab-ranks, are specified in advance and the heuristic chooses the cab-rank where the idle vehicle should be dispatched. To take a decision, the heuristic exploits information about future patterns of demand at each cab-rank.

Finally, Coslovich et al. (2003) have addressed a dial-a-ride problem where people might unexpectedly ask a driver for a trip at a given stop. Clearly, these requests must immediately be accepted or rejected. In order to accommodate them, a neighborhood of solutions is generated off-line by considering different perturbations to the current planned routes. Using this neighborhood of solutions, more insertion opportunities are available and can be quickly evaluated when an unexpected customer asks for service. Whenever a new customer is accepted, both the current solution and the neighborhood must be updated, but this computationally demanding task can be done while the vehicles are moving from one stop to the next. In that case, the time pressure is much less stringent, when compared with the almost immediate response required by unexpected customers.

4 Urban courier service problems

Every large city has a number of courier companies serving pickup and delivery requests for the transportation of letters and small parcels. These requests occur continuously during the day and only a small fraction of these are known in advance (typically, those requests that have been received the previous day, but too late for immediate service). The distinctive features of Urban Courier Service Problems (UCSPs) with regard to the other problems presented in this chapter are their inherent dynamic nature and the absence of capacity constraints, due to the small size of letters and parcels.

In these problems, each request is characterized by a pickup and a delivery location, plus a time window for service. A usually fixed fleet of vehicles is available to service the requests. When a new request occurs, it is dispatched and inserted in a least-cost fashion in the planned route of one vehicle. This cost typically relates to the total distance traveled by the vehicles plus a penalty for lateness when the vehicle arrives at a location after its time window's upper bound (a vehicle can arrive before the lower bound but, in that case, it must wait up to the lower bound to start its service). With a fixed size fleet, some requests received during the day may remain unserviced. This happens, for example, when drivers must be back at some location before a given deadline at the end of the day, when the upper bounds of the time windows are strictly enforced, or when the incremental cost for servicing a request exceeds a tolerance threshold. In practice, these unserviced requests will be serviced the next day or will be referred to alternative transportation means, including competitors.

This type of problem has not been studied much in the literature. In particular, we are not aware of any exact methods for solving them and only a few

heuristics are reported. Furthermore, the only dynamic aspect of the problem that has been considered is the occurrence of new requests, although other aspects are certainly of interest like dynamic travel times. In the following, the main algorithms for the UCSP are reviewed.

The single-vehicle dynamic pickup and delivery problem (without time windows) was analyzed by Swihart and Papastravou (1999) under different routing policies and demand intensities. This paper can be seen as an extension of the work of Bertsimas and van Ryzin (1991) on the Dynamic Traveling Repairman Problem (DTRP) with single-point customer requests. As the objective is to minimize the time spent in the system, this work is at the interface of vehicle routing and queuing theory. Dynamic contexts with both unit-capacity and multiple-capacity vehicles are analyzed.

4.1 A tabu search heuristic for the UCSP

A natural heuristic approach for the UCSP is to use a cheapest insertion criterion for new requests. Gendreau et al. (1998) go beyond this principle and reoptimize the planned routes with a tabu search heuristic. The neighborhood structure is based on ejection chains (Glover, 1996), where the pickup and delivery locations of a request are taken from one route and moved to another route, forcing a request from that route to move to yet another route, and so on. The chain may be of any length and may be cyclic or not. The "best" ejection chain is obtained by solving a constrained shortest path problem. The tabu search heuristic also integrates an adaptive memory (Rochat and Taillard, 1995), which combines high quality solutions to produce new ones.

Due to real-time requirements, different computational techniques are proposed to speed up the neighborhood evaluation. In addition, a parallel implementation is developed where a master processor distributes the workload among the slaves that run the tabu search processes. A two-level parallelization scheme is used. First, several tabu search threads run in parallel and communicate through the common adaptive memory: they all feed the memory with new improved solutions and get starting solutions from it. Second, each solution (a set of planned routes) within a search thread is partitioned into subsets of routes, and a different tabu search process is associated with each subset. This is a form of intensification, as it allows the search to focus on restricted parts of the solution, while reducing the computational effort as a side effect.

The tabu search heuristic runs between the occurrence of new events. At each input update, it solves the static problem associated with known requests. Because it does not account for future requests, it can be characterized as a "myopic" problem-solving approach. Two different types of event are considered: the occurrence of new service requests, which are truly dynamic events, and the completion of service at customer locations (which can be determined from the current routes, due to deterministic travel times). When a new request is received, the tabu search processes are stopped and their best solution is sent to the master for possible inclusion in the adaptive memory. The new

request is then inserted at least cost in each solution contained in the adaptive memory. Once updated, the memory feeds the tabu search processes with new starting solutions. A similar procedure is applied when service is completed at a given location. In that case, the best solution in the adaptive memory is used to identify the vehicle's next destination, and the other solutions in the memory are updated accordingly.

A simulator was developed to test the algorithm under realistic scenarios with up to 30 requests per hour. The computational results demonstrated the superiority of the tabu search heuristic for handling new requests, when compared with more straightforward approaches, like simple insertion heuristics. It is thus useful to optimize the planned routes with sophisticated procedures, even when the optimization does not account for future requests. Although this is not explicitly mentioned by Gendreau et al. (1998), the benefits mostly arise from the early portion of the planned route (as opposed to the later portion which is likely to be modified with the arrival of new requests). This observation naturally leads to the work described in the next subsection.

4.2 A double-horizon strategy for the UCSP

Mitrović-Minić et al. (2004) have proposed a double-horizon strategy for solving a variant of the UCSP. In this work, the number of vehicles is a free variable, thus allowing all requests to be serviced. Furthermore, each request must be visited before a deadline, which means that the objective reduces to minimizing the total distance traveled. The authors propose a generalization of the short-term rolling horizon approach (where only requests with a time window sufficiently close to the current time are assigned to planned routes (Psaraftis, 1988)). Both a short-term and a long-term planning horizon are considered. The latter is introduced to alleviate the adverse long-term effects of apparently good short-term decisions. Basically, the idea is to associate a different objective with each horizon type. The objective associated with the short term horizon is the true objective (i.e., total distance traveled), while the objective associated with the long-term horizon is aimed at introducing large waiting times in the routes to favor the insertion of future requests. Each objective is optimized with a simplified version of the tabu search heuristic reported in Section 4.1. The computational results obtained on instances generated from data collected from two courier companies, operating in Vancouver, Canada, demonstrate the benefits of the double-horizon approach when compared to a single horizon approach. Mitrović-Minić and Laporte (2004) have further analyzed different ways of introducing waiting times when scheduling the planned routes. They showed that mixed waiting strategies provide better results. The best approach partitions a planned route into segments made of close locations. Within a segment, the vehicle always departs as soon as possible from its current location; but when it is time to cross a boundary between two segments to travel further, the vehicle waits at its current location for a fraction of the time available up to the latest possible departure time.

4.3 Adaptive methods for the UCSP

The approaches reported in this subsection exploit the knowledge accumulated, over the years, by expert human dispatchers to make good dispatching decisions. They can be considered as learning or adaptive methods. In the work of Shen et al. (1995), a neural network model learns to assign requests to vehicles by automatically adjusting itself to a sample of decisions previously made by an expert (see Rumelhart et al., 1986, for an introduction to feedforward neural networks and the backpropagation learning algorithm). After training with 140 dispatching scenarios obtained from a small courier company operating in Montreal, and for which expert decisions were known, the network was able to take good dispatching decisions on other sets of previously unseen scenarios.

Leclerc and Potvin (1997) proposed a linear utility function that integrates the main decision variables considered by expert dispatchers when they make decisions. The decision variables within the utility function are then weighted with a genetic algorithm (Holland, 1992). Basically, different sets of weights "evolve" through genetic mechanisms, with the objective of matching as closely as possible a sample of decisions previously taken by an expert. Benyahia and Potvin (1998) extended this work by evolving nonlinear utility functions, using a genetic programming framework (Koza, 1992).

5 The dial-a-flight problem

Almost 3000 businesses in the United States provide on-demand air charter services (certificated by the FAA as Part 135 on-demand air charter). The majority of companies in the industry are small businesses regulated by the FAA with similar oversight to that given to the large scheduled airlines. The on-demand air charter industry provides a vital transportation link for medical services, important cargo needed to promote commerce, and personal travel supporting the growth of the economy. These companies use smaller aircraft to meet the customized needs of the traveling public for greater flexibility in scheduling and access to almost every airport in the country. Flights are planned according to the customer's schedule, not the operator's. On-demand air charter serves commerce across the country and the world by providing short notice delivery of parts, important documents, supplies, and other valuable cargo. On-demand air charter also saves lives, since air ambulances transport critically ill or injured patients to hospitals and trauma centers that can provide the necessary care, and transport vital organs for those requiring transplants. All of these services are contingent upon the ability to respond quickly to the needs of customers.

Recently there has been an increased interest in the passenger service sector of the on-demand air charter industry. This is, in part, due to the changes taking place in the commercial airline industry. Increased security at airports

has resulted in longer waiting times, with the associated frustrations, and thus longer travel times. Furthermore, due to the huge losses suffered by the airlines in recent years (airlines worldwide have lost $25 billion and more than 400,000 jobs in 2002 and 2003), airlines have cut back and are operating with a reduced schedule, affecting the flexibility of the business traveler, especially when it concerns smaller regional airports. At the same time, technological advances are paving the way for the development of cheaper jet airplanes. For example, the Eclipse 500, a six-seat, single-pilot state-of-the art jet will sell for about $1 million – about one-quarter of the price of the cheapest business jets made today. As a consequence, the idea of an *air taxi* service, providing efficient, hassle-free, affordable, on-demand air transportation, is no longer just fiction; it is rapidly becoming a reality. In fact, an air taxi already exists today. Since April 2002, SkyTaxi, Inc. (*http://www.skytaxi.com*) has been providing on-demand air transportation in the northwestern United States. Some passengers pay a premium to fly direct and by themselves; others receive a discount by agreeing to allow stops to pick up or drop off other passengers.

There are obvious advantages to an air taxi system. More than 1.5 million people board commercial airliners each day. Most fly with hundreds of other passengers on jumbo jet airplanes to and from a limited number of major "hub" airports which are heavily congested and often located many miles from their homes and final destinations. Missed connections and flight delays add to their frustrations. An air taxi system gives travelers the option of hopping aboard small jets that fly to and from less congested outlying airports, without packed parking lots, long lines at security checkpoints, flight delays, and lost luggage, that are closer to where they live and where they want to go. While commercial flight service now exists at only about 550 airports in the United States, air taxis will be able to land at 10,000 of the nation's 14,000 public and private runways. And all that at competitive fares.

Even though many characteristics of an air taxi service are similar to those of the common taxi cab service, there are also some fundamental differences. Requests for service are typically placed farther in advance, a day to two days in advance as opposed to minutes to an hour. This gives more time to optimize a flight schedule. Usually, a request for service involves several origin-destination pairs, because business travelers in the end want to return home. Furthermore, the set of possible pickup, drop-off, and transit locations is relatively small and known in advance. An air taxi service operates out of a given set of airports, which implies a flight schedule optimizer can exploit the fixed transportation network structure. Finally, the FAA imposes strict rules on pilot flying and duty hours as well as on aircraft maintenance, which all affect scheduling flexibility.

Consequently, the traditional dial-a-ride problem is insufficient as an abstract representation for studying, analyzing, and developing decision technology for air taxi services. In this section, we introduce the Dial-a-Flight Problem (DAFP). As air taxi services are a new phenomenon, there is little or no lit-

erature on dial-a-flight problems, so we will focus on problem definition and optimization challenges.

5.1 Problem definition

A key challenge in setting up an air taxi system is the development of a scheduling engine that takes requests for transportation and schedules planes and pilots in a cost effective way to satisfy these requests. The static and dynamic DAFPs defined below capture the essential characteristics of the scheduling problems encountered by an air taxi service.

5.1.1 The static dial-a-flight problem

The Static Dial-a-Flight Problem (SDAFP) is concerned with the scheduling of a set P of n single passenger requests for air transportation during a single day. A request $i \in P$ specifies an origin airport o_i, an earliest acceptable departure time e_i at o_i, a destination airport d_i, and a latest acceptable arrival time l_i at d_i. A request i results in a revenue of r_i. A fleet F of m identical airplanes with capacity Q and operable by a single pilot is available to provide the requested air transportation. Each airplane $j \in F$ has a home base B_j, is available between E_j and L_j, and returns to its home base at the end of the day. A set P of pilots, stationed at the home bases of the airplanes, are available to fly the airplanes. A pilot departs from the home base at the start of his duty and returns to his home base at the end of his duty. A pilot schedule has to satisfy FAA regulations governing flying hours and duty period, i.e., a pilot cannot fly more than 8 hours in a day and his duty period cannot be more than 14 hours. It takes an airplane t_{uv} time to fly from airport u to v (over a distance d_{uv}) and a cost c_{uv} is incurred when doing so. To ensure acceptable service a passenger itinerary will involve at most two flights, i.e., a single intermediate stop is allowed. The turnaround time at an airport, i.e., the minimum time between an arrival at an airport and the next departure, is 30 minutes. The objective is to maximize the profit, i.e., revenues minus costs, while satisfying all requests. (Note that because all requests have to be satisfied, in this variant of the problem the revenues are fixed and the objective is to minimize the costs.) A dispatcher has to decide which planes and pilots to use to satisfy the requests and what the plane and pilots itineraries will be, i.e., the flight legs and associated departure times.

Several restrictions in the above problem definition represent business decisions rather than physical limitations. For example, the limit on the duration of a passenger trip from origin to destination is a business rule, set, for example, to twice the direct flying time. This constraint which is similar to the maximum ride time in the DARP reflects a trade-off between scheduling flexibility and customer service. Intuitively, an important factor contributing to profitability for an air taxi service provider is a high plane utilization. This can be accomplished by having an airplane satisfy multiple requests at the same time. Therefore, an air taxi service provider may "encourage" passengers to accept a

trip that involves an intermediate stop (to pickup or drop off other passengers). Note that the turnaround time also affects flexibility, and therefore profitability. Another business rule, which has not been made explicit in the problem definition, is whether or not to allow passengers to change planes at an intermediate stop. Again, allowing this will increase the scheduling flexibility, but it may have a negative impact on customer service because of potential delays.

5.1.2 The dynamic dial-a-flight problem

In the Dynamic Dial-a-Flight Problem (DDAFP) the set of requests for air transportation arrives over time and each time a request arrives one must immediately decide whether it is feasible to accept the request given the available resources and the commitments already made. In addition, if it is feasible to accept the request, one may also want to decide whether it is desirable to accept the request, i.e., whether it will increase profit. The latter decision is especially complex as it depends on the requests that will arrive in the future.

The simplest variant of the DDAFP is to construct a schedule of flights for a specific day in the not too distant future. Requests for transportation on that particular day arrive in real-time and are considered up to a certain cut-off time, which precedes the actual execution of the planned schedule. The rule is to accept each request if there is available capacity. In the DDAFP, one needs to accommodate bundles of requests, as customers typically request not one, but two or more flights, and ultimately want to return to their point of origin. Clearly, all these flight requests must be accepted as a bundle, rather than individually. Note that all requests are received prior to the execution of any flight schedule. A more complex variant incorporates "same day travel" service, where requests can arrive during the execution of a flight schedule and have to be incorporated into the schedule.

The real-time, online nature of the booking process is not specific to the air taxi service, but a common feature of many transportation problems. The literature on dynamic and stochastic routing problems is rapidly expanding. For a discussion of many of the issues related to accept/reject decisions in transportation problems, see Bent and Van Hentenryck (2004) and Campbell and Savelsbergh (2003, 2005).

5.2 Solution approaches for dial-a-flight problems

As mentioned earlier, little or no literature exists on dial-a-flight problems. Therefore, we present a natural integer programming formulation for SDAFP and we discuss some pertinent issues concerning solution approaches for the DDAFP.

5.2.1 A model for the static dial-a-flight problem

In this subsection, we present a time-discretized multicommodity network flow model for one variant of the problem. We assume the company providing

the air taxi service has decided to operate each of its planes with two pilot shifts per day, a morning shift and an afternoon shift. The plane has to return to its home base to switch pilots some time in the early afternoon. The shift lengths are chosen so that the limit on pilot duty hours is automatically satisfied. Finally, it is assumed that for the expected demand distribution it is unlikely that the limit on the number of flying hours of the pilots is violated, and therefore, pilot constraints are not explicitly incorporated in the model.

The time horizon is discretized into time periods of several minutes. Let T be the set of time periods in the planning horizon and let A be the set of airports. Define the following decision variables:

$$x^i_{uvt} = \begin{cases} 1 & \text{if passenger } i \text{ departs from airport } u \text{ to airport } v \\ & \quad \text{at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

$$y^j_{uvt} = \begin{cases} 1 & \text{if airplane } j \text{ departs from airport } u \text{ to airport } v \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

For each airplane j, let R_j and S_j define the start and end of the period during which the pilot switch needs to take place.

The mathematical formulation of SDAFP can now be written as follows (where parameters have been converted to their time period equivalents whenever necessary):

$$\text{minimize} \quad \sum_j \sum_u \sum_{u \neq v} \sum_t c_{uv} y^j_{uvt}$$

subject to

$$\sum_v x^i_{uvt} = 1, \quad i \in P, u = o_i, t = e_i, \tag{35}$$

$$\sum_v x^i_{vu(t-t_{vu})} = 1, \quad i \in P, u = d_i, t = l_i, \tag{36}$$

$$x^i_{uu(t-1)} + \sum_{v \neq u} x^i_{vu(t-t_{vu})} - \sum_{v \neq u} x^i_{uvt} - x^i_{uut} = 0,$$
$$i \in P, u \in A, t = e_i + 1, \ldots, l_i - 1, \tag{37}$$

$$\sum_u \sum_{v \neq u} \sum_t x^i_{uvt} \leqslant 2, \quad i \in P, \tag{38}$$

$$\sum_i x^i_{uvt} - \sum_j Q y^j_{uvt} \leqslant 0, \quad u \in A, v \in A, u \neq v, t \in T, \tag{39}$$

$$\sum_v y^j_{uvt} = 1, \quad j \in F, u = B_j, t = E_j, \tag{40}$$

$$\sum_v y^j_{vu(t-t_{vu})} = 1, \quad j \in F, u = B_j, t = L_j, \tag{41}$$

$$y^j_{uu(t-1)} + \sum_{v \neq u} y^j_{vu(t-t_{vu})} - \sum_{v \neq u} y^j_{uvt} - y^j_{uut} = 0,$$
$$j \in F, u \in A, t = E_j + 1, \ldots, L_j - 1, \tag{42}$$
$$\sum_{v} \sum_{t=R_j - t_{vu}}^{t=S_j - t_{vu}} y^j_{vut} \geqslant 1, \quad j \in F, u = B_j, \tag{43}$$
$$x^i_{uvt} \in \{0, 1\}, \quad i \in P, u \in A, v \in A, t \in T, \tag{44}$$
$$y^j_{uvt} \in \{0, 1\}, \quad j \in F, u \in A, v \in A, t \in T. \tag{45}$$

Constraints (35) and (36) state that each passenger departs from his origin airport and arrives at his destination airport. Constraints (37) ensure passenger flow conservation at airports. Similar constraints are specified for each airplane, i.e., (40)–(42). Constraints (38) limit the number of intermediate stops for each passenger and constraints (39) enforce the airplane capacity, i.e., the total number of passengers flying on a leg is less than the total capacity of all airplanes flying on that leg. Finally, constraints (43) force each airplane to return to its home base to switch pilots.

This time-discretized multicommodity network flow model becomes large quickly and specialized solution approaches need to be developed to solve even medium size instances, e.g., involving 15–30 airports and 5–10 airplanes.

5.2.2 Algorithmic issues related to the dynamic dial-a-flight problem

In the DDAFP one has to decide, given a set of already accepted requests, whether an incoming request can be served or not. The amount of time available to make this decision depends on the business rules but most likely there is very little time to do so. Consequently, fast heuristics will have to be part of the decision technology. It is, however, conceivable that when heuristics fail to accept a request quickly, a customer may be given the option of receiving final notification of acceptance or rejection in, say, 30 minutes to allow time for optimization based techniques to try and accommodate the request. Research on the DDAFP is relatively new and, in the current state of knowledge, there are more questions than answers.

Given a set of already accepted requests and an existing feasible schedule, it makes sense to ask what is the value of that schedule in deciding whether it is feasible to accept a new request. What if simple insertion into any of the existing airplane itineraries is not feasible? Does one try to reoptimize a single airplane itinerary or try to reoptimize several airplane itineraries? One by one or simultaneously? How does one select the airplane itineraries to consider? What if the existing schedule included passenger plane changes? In fact, there is no need to keep just one feasible schedule for the already accepted requests. Should one keep several feasible schedules and try to insert new requests in each of them?

Once an incoming request has been accepted, there may be time available to optimize the schedule of the accepted requests by solving a static dial-a-flight

problem. Does that help handle the next request? Should the standard objective function be used or is it better to use an objective function that focuses on the remaining flexibility in the schedule? How does one formally define remaining flexibility?

5.3 The dial-a-flight problem in practice

The DAFP is a complex and challenging optimization problem, but still ignores many relevant practical aspects. For example, in reality the costs of a flight leg is a function of the fuel consumption, which depends on the weight of the plane, the altitude of the flight, and the duration of the flight (shorter flights burn more fuel as most fuel is consumed during take off and landing). Furthermore, fuel prices may differ at various locations, so deciding where to refuel may impact overall costs. Also, taxi service providers will likely offer various service classes, e.g., for a higher price a direct flight will be guaranteed, for an even higher price one can charter an entire airplane plus a pilot for certain period of time.

6 Ambulance fleet management

Transportation on demand problems also arise in the planning of emergency services associated with fire fighting, police patrols, and ambulance fleet planning. Several important location, staffing and dispatching issues lie at the heart of the problems encountered in these three areas but ambulance fleet management is certainly the most relevant to this chapter. Readers interested in the management of fire companies and in police patrols are referred to the work of Larson (1972), Walker et al. (1979), Swersey (1994), and Adams (1997).

One of the key issues arising in ambulance fleet management is to decide where to locate ambulances to provide at all times an adequate population coverage. Over the past thirty-five years, there has been a steady evolution in the development of ambulance location models, as witnessed by the work of Marianov and ReVelle (1995) and Brotcorne et al. (2003). The latter study serves as a basis for this chapter. The first models developed in the 1970s proposed static solutions to a problem that is essentially stochastic and dynamic. Over the years more realism has gradually been introduced into these basic models until the emergence, a few years ago, of a truly dynamic solution procedure.

Most contributions in the field of ambulance location present integer linear programming formulations, but no solution techniques. Only in recent years have algorithms been proposed.

6.1 Two early ambulance location models

Ambulance location models are usually defined on a graph $G = (V \cup W, A)$, where V is a node set representing aggregated demand points, W is a set of

potential ambulance location sites, and $A = \{(i, j) \in V \cup W\}$ is an arc set. With each arc (i, j) is associated a travel time t_{ij}. A demand point $i \in V$ is *covered* by site $j \in W$ if and only if $t_{ij} \leqslant r$, where r is a preset coverage standard. Let $W_i = \{j \in W\colon\ t_{ij} \leqslant r\}$ be the set of location sites covering demand point i.

The Location Set Covering Model (LSCM) of Toregas et al. (1971) aims to minimize the number of ambulances needed to cover all demand points. It uses binary variables x_j equal to 1 if and only if an ambulance is located at j:

$$\text{(LSCM)} \quad \text{minimize} \quad \sum_{j \in W} x_j \tag{46}$$

subject to

$$\sum_{j \in W_i} x_j \geqslant 1, \quad i \in V, \tag{47}$$

$$x_j \in \{0, 1\}, \quad j \in W. \tag{48}$$

The Maximum Covering Location Problem (MCLP) of Church and ReVelle (1974) works with a fixed number p of ambulances and attempts to cover the largest possible demand $z(p)$. Denote by d_i the demand at node $i \in V$ and let y_i be a binary variable equal to 1 if and only if i is covered by at least one ambulance:

$$\text{(MCLP)} \quad \text{maximize} \quad z(p) = \sum_{i \in V} d_i y_i \tag{49}$$

subject to

$$\sum_{j \in W_i} x_j \geqslant y_i, \quad i \in V, \tag{50}$$

$$\sum_{j \in W} x_j = p, \tag{51}$$

$$x_j \in \{0, 1\}, \quad j \in W, \tag{52}$$

$$y_i \in \{0, 1\}, \quad i \in V. \tag{53}$$

A good way to combine these two models is to repeatedly solve MCLP with increasing values of p and select a solution offering a good compromise between p and $z(p)$.

6.2 *Static models with extra coverage*

One major drawback of LSCM and of MCLP is that adequate coverage may no longer exist once an ambulance has been dispatched. This is the reason why models with extra coverage have been introduced. The Tandem Equipment Allocation Model (TEAM) of Schilling et al. (1979) works with two equipment types A and B, corresponding to advanced life support (ALS) units and basic life support units (BLS) operating with different time standards (see Mandell,

1998). Let r^A and r^B be the respective coverage standards of A and B, and let $W_i^A = \{j \in W\colon t_{ij} \leqslant r^A\}$ and $W_i^B = \{j \in W\colon t_{ij} \leqslant r^B\}$. Let x_j^A (resp. x_j^B) be a binary variable equal to 1 if and only if a vehicle of type A (resp. B) is located at $j \in W$, and let y_i be a binary variable equal to 1 if and only if $i \in V$ is covered by two types of vehicle.

$$\text{(TEAM)} \quad \text{maximize} \quad \sum_{i \in V} d_i y_i \tag{54}$$

subject to

$$\sum_{j \in W_i^A} x_j^A \geqslant y_i, \quad i \in V, \tag{55}$$

$$\sum_{j \in W_i^B} x_j^B \geqslant y_i, \quad i \in V, \tag{56}$$

$$\sum_{j \in W} x_j^A = p^A, \tag{57}$$

$$\sum_{j \in W} x_j^B = p^B, \tag{58}$$

$$x_j^A \leqslant x_j^B, \quad j \in W, \tag{59}$$

$$x_j^A, x_j^B \in \{0, 1\}, \quad j \in W, \tag{60}$$

$$y_i \in \{0, 1\}, \quad i \in V. \tag{61}$$

This model is a direct extension of MCLP, with the proviso that constraints (59) impose a hierarchy between the two vehicle types. In the FLEET model of Schilling et al. (1979), these constraints are relaxed and the number of potential location sites is limited to a preset value p. Another variant, proposed by Daskin and Stern (1981), is to use a hierarchical objective to first maximize the number of demand points covered more than once.

In the same spirit, Hogan and ReVelle (1986) have proposed the Backup Coverage Problems, called BACOP1 and BACOP2, in which x_j is the number of ambulances located at $j \in W$, and y_i, u_i are binary variables equal to 1 if and only if $i \in V$ is covered once or at least twice, respectively:

$$\text{(BACOP1)} \quad \text{maximize} \quad \sum_{i \in V} d_i u_i \tag{62}$$

subject to

$$\sum_{j \in W_i} x_j \geqslant 1 + u_i, \quad i \in V, \tag{63}$$

$$\sum_{j \in W} x_j = p, \tag{64}$$

$$u_i \in \{0, 1\}, \quad i \in V, \tag{65}$$

$$x_j \geqslant 0 \text{ and integer}, \quad j \in V, \tag{66}$$

and

$$\text{(BACOP2)} \quad \text{maximize} \quad \theta \sum_{i \in V} d_i y_i + (1 - \theta) \sum_{i \in V} d_i u_i \tag{67}$$

subject to

$$\sum_{j \in W_i} x_j \geqslant y_i + u_i, \quad i \in V, \tag{68}$$

$$u_i \leqslant y_i, \quad i \in V, \tag{69}$$

$$\sum_{j \in W} x_j = p, \tag{70}$$

$$u_i \in \{0, 1\}, \quad i \in V, \tag{71}$$

$$y_i \in \{0, 1\}, \quad i \in V, \tag{72}$$

$$x_j \geqslant 0 \text{ and integer}, \quad j \in W, \tag{73}$$

where θ is a weight chosen in $[0, 1]$.

The Double Standard Model (DSM) of Gendreau et al. (1997) works with two coverage standards r_1 and r_2, with $r_1 < r_2$, as specified by the United States Emergency Medical Services Act of 1973. A proportion α of the demand must be covered within r_1 while the entire demand must be covered within r_2. In the DSM, the objective is to maximize the demand covered twice within r_1 using p ambulances, at most p_j of which are located at $j \in W$, subject to the double coverage constraints. Let $W_i^1 = \{j \in W\colon\ t_{ij} \leqslant r_1\}$ and $W_i^2 = \{j \in W\colon\ t_{ij} \leqslant r_2\}$. The integer variable x_j denotes the number of ambulances located at $j \in W$ and the binary variable y_i^k is equal to 1 if and only if the demand at node $i \in V$ is covered k times ($k = 1$ or 2) within r_1. The formulation is then:

$$\text{(DSM)} \quad \text{maximize} \quad \sum_{i \in V} d_i y_i^2 \tag{74}$$

subject to

$$\sum_{j \in W_i^2} x_j \geqslant 1, \quad i \in V, \tag{75}$$

$$\sum_{i \in V} d_i y_i^1 \geqslant \alpha \sum_{i \in V} d_i, \tag{76}$$

$$\sum_{j \in W_i^1} x_j \geqslant y_i^1 + y_i^2, \quad i \in V, \tag{77}$$

$$y_i^2 \leqslant y_i^1, \quad i \in V \tag{78}$$

$$\sum_{j \in W} x_j = p, \tag{79}$$

$$x_j \leqslant p_j, \quad j \in W, \tag{80}$$

$$y_i^1, y_i^2 \in \{0, 1\}, \quad i \in V, \tag{81}$$

$$x_j \geqslant 0 \text{ and integer}, \quad j \in W. \tag{82}$$

Here, the objective function computes the demand covered twice within r_1 time units, and constraints (75) mean that all demand is covered within r_2. The left-hand side of (77) represents the number of ambulances covering node i within r_1 units, while the right-hand side is equal to 1 if i is covered once within r_1 units, and equal to 2 if it is covered at least twice within r_1 units. The combination of constraints (76) and (77) ensures that a proportion α of the demand is covered within r_1. Constraints (78) state that node i cannot be covered at least twice if it is not covered at least once. In constraints (80), p_j can be set equal to 2 since an optimal solution using this upper bound always exists.

Gendreau, Laporte, and Semet solve the DSM by means of a tabu search procedure in which neighbor solutions are obtained by generating sequences of ambulance moves to an adjacent location, not unlike what is done in ejection chain methods (see, e.g., Rego and Roucairol, 1996). Comparisons with the linear relaxation value of the model indicate that the heuristic typically yields optimal or near-optimal solutions.

6.3 *Probabilistic models with extra coverage*

None of the models introduced so far takes into account that ambulances are not always available to answer a call. A way around this is to assume that each ambulance has a probability q, called *busy fraction*, of being unavailable. This value is obtained by dividing the total time spent by all ambulances on all calls by the total ambulance time available. If $i \in V$ is covered by k ambulances, then the expected demand covered at that node is $E_{i,k} = d_i(1 - q^k)$ and the marginal contribution of the kth ambulance is $E_{i,k} - E_{i,k-1} = d_i(1 - q)q^{k-1}$.

In the Maximum Expected Covering Location Model (MEXCLP) of Daskin (1983), up to p ambulances may be located in total, and more than one vehicle may be located at the same node. Let y_{ik} be a binary variable equal to 1 if and only if node $i \in V$ is covered by at least k ambulances. The model can be written as follows:

$$\text{(MEXCLP)} \quad \text{maximize} \quad \sum_{i \in V} \sum_{k=1}^{p} d_i(1 - q)q^{k-1} y_{ik} \tag{83}$$

subject to

$$\sum_{j \in W_i} x_j \geqslant \sum_{k=1}^{p} y_{ik}, \quad i \in V, \tag{84}$$

$$\sum_{j \in W} x_j \leqslant p, \tag{85}$$

$$x_j \geqslant 0 \text{ and integer}, \quad j \in W, \tag{86}$$

$$y_{ik} \in \{0, 1\}, \quad i \in V, k = 1, \ldots, p. \tag{87}$$

The validity of this model stems from the fact that the objective function is concave in k. Therefore, if $y_{ik} = 1$, then $y_{ih} = 1$ for $h \leqslant k$. Since the objective is to be maximized, both (84) and (85) will be satisfied as equalities. It follows that the two sides of (84) will be equal to the number of ambulances covering node $i \in V$.

An application of MEXCLP to the City of Bangkok ($|V| = 59$, $|W| = 46$, $10 \leqslant p \leqslant 30$) by Fujiwara et al. (1987) has shown that without reducing expected coverage, the number of ambulances could be reduced to 15 from the current 21. Repede and Bernardo (1994) have proposed TIMEXCLP, a dynamic implementation of MEXCLP, in which travel speeds are allowed to vary during the day. Goldberg et al. (1990) have worked with stochastic travel times. They compute the probability that a given demand point i will be covered based on three probabilities: the probability that an ambulance located at the sth preferred site of i can reach i within eight minutes, the probability that this ambulance is available, and the probability that ambulances located at less preferred sites are not available. Experiments conducted on data collected in Tucson, Arizona, have shown that the use of this model could increase the expected covered demand from 24% to 53.1%.

ReVelle and Hogan (1989) have developed two chance-constrained maximal covering location models, called the Maximal Availability Location Problem (MALP I and MALP II), in which the constraint

$$1 - q^{\sum_{j \in W_i} x_j} \geqslant \alpha, \quad i \in V, \tag{88}$$

ensures that each demand point is covered with probability at least equal to α. This constraint, which can be linearized as

$$\sum_{j \in W_i} x_j \geqslant \left\lceil \frac{\log(1 - \alpha)}{\log q} \right\rceil = b, \quad i \in V, \tag{89}$$

can be used instead of (47) in LSCM. In MALP I, Hogan and ReVelle maximize the total demand covered with b ambulances, subject to a global availability of p ambulances. Defining binary variables y_{jk} as in MEXCLP, the MALP I model can then be written as:

$$\text{(MALP I)} \quad \text{maximize} \quad \sum_{i \in V} d_i y_{ib} \tag{90}$$

subject to

$$\sum_{k=1}^{b} y_{jk} \leqslant \sum_{j \in W_i} x_j, \quad i \in V, \tag{91}$$

$$y_{ik} \leqslant y_{i,k-1}, \quad i \in V, k = 2, \ldots, b, \tag{92}$$

$$\sum_{j \in W} x_j = p, \tag{93}$$

$$x_j \in \{0, 1\}, \quad j \in W, \tag{94}$$

$$y_{ik} \in \{0, 1\}, \quad i \in V, k = 1, \ldots, p. \tag{95}$$

Here, constraints (92) are required since the concavity property observed in MEXCLP no longer holds.

In MALP II, ReVelle and Hogan estimate a busy fraction q_i associated with each i, as the ratio of the total duration of all calls associated to i to the total ambulance time in W_i. The major difficulty with this is that q_i cannot be computed a priori because it is an output of the model. An iterative procedure is then required to solve MALP II approximately. Finally we mention the existence of two studies, by Batta et al. (1989) and by Marianov and ReVelle (1994), whose aim is to better approximate the busy fraction of the whole system or associated with a particular location. These are based in part on Larson's (1974) hypercube model.

Finally, Ball and Lin (1993) have developed an extension of LSCM, called *Rel-P*, which incorporates a linear constraint on the number of vehicles required to achieve a given reliability level. The model contains binary variables x_{jk} equal to 1 if and only if k ambulances are located at node $j \in W$, and constants c_{jk} equal to the cost of locating k vehicles at site j. An upper bound p_j is imposed on the number of ambulances located at site j. Their model is as follows:

$$\text{(Rel-P)} \quad \text{minimize} \quad \sum_{j \in J} \sum_{1 \leqslant k \leqslant p_j} c_{jk} x_{jk} \tag{96}$$

subject to

$$\sum_{1 \leqslant k \leqslant p_j} x_{jk} \leqslant 1, \quad j \in W, \tag{97}$$

$$\sum_{j \in W_i} \sum_{1 \leqslant k \leqslant p_j} a_{jk} x_{jk} \geqslant b_i, \quad i \in V, \tag{98}$$

$$x_{jk} \in \{0, 1\}, \quad j \in W, 1 \leqslant k \leqslant p_j. \tag{99}$$

In constraints (98), the constants a_{jk} and b_i are computed to ensure that given the number of ambulances covering demand point i, the probability of being unable to answer a call does not exceed a certain value. The computation of the a_{jk} and b_i coefficients are, in fact, carried out by using an upper bound on that probability.

Ball and Lin incorporate valid inequalities in their model which is then solved by means of a standard branch-and-bound code for integer linear programming.

6.4 A dynamic model

In practice, ambulances are often relocated over time in order to always ensure an adequate coverage. Gendreau et al. (2001) have developed a dynamic relocation model where a new redeployment can in principle be implemented at each instant t at which an ambulance is dispatched to a call or returns from a call. The model is based on the DSM developed by the same authors (Gendreau et al., 1997). It constitutes, to our knowledge, the only available dynamic ambulance relocation tool. In addition to the standard coverage and site capacity constraints, the model takes into account a number of practical considerations inherent to the dynamic nature of the problem: (1) one should avoid relocating the same vehicle frequently within a short time interval; (2) repeated round trips between the same two location sites must be avoided; (3) long trips between the initial and final location sites must be avoided.

The dynamic aspect of the redeployment model is captured by time dependent constants $M^t_{j\ell}$ equal to the cost of relocating, at time t, ambulance ℓ from its current site to site $j \in W$. This includes the case where site j coincides with the current location of the ambulance, i.e., $M^t_{j\ell} = 0$. The constant $M^t_{j\ell}$ captures some of the history of ambulance ℓ. If it has been relocated frequently prior to time t, then $M^t_{j\ell}$ will be larger. If relocating ambulance ℓ to site j violates any of the above constraints, then the relocation is simply disallowed. The constant p^t is the number of ambulances available at time t, and p^t_j is the maximum number of ambulances that can be located at site j at time t. Binary variables $x_{j\ell}$ are equal to 1 if and only if ambulance ℓ is relocated to site j. The Dynamic Double Standard Model at Time t (DDSM^t) can now be described:

$$(\text{DDSM}^t) \quad \text{maximize} \quad \sum_{i\in V} d_i y^2_i - \sum_{j\in W}\sum_{\ell=1}^{p^t} M^t_{j\ell} x_{j\ell} \tag{100}$$

subject to

$$\sum_{j\in W^2_i}\sum_{\ell=1}^{p^t} x_{j\ell} \geqslant 1, \quad i \in V, \tag{101}$$

$$\sum_{i\in V} d_i y^1_i \geqslant \alpha \sum_{i\in V} d_i, \tag{102}$$

$$\sum_{j\in W^1_i}\sum_{\ell=1}^{p^t} x_{j\ell} \geqslant y^1_i + y^2_i, \quad i \in V, \tag{103}$$

$$y^2_i \leqslant y^1_i, \quad i \in V, \tag{104}$$

$$\sum_{j\in W} x_{j\ell} = 1, \quad \ell = 1, \ldots, p^t, \tag{105}$$

$$\sum_{\ell=1}^{p^t} x_{j\ell} \leqslant p_j^t, \quad j \in W, \tag{106}$$

$$y_i^1, y_i^2 \in \{0, 1\}, \quad i \in V, \tag{107}$$

$$x_{j\ell} \in \{0, 1\}, \quad j \in W, \ell = 1, \ldots, p^t. \tag{108}$$

Apart from variables $x_{j\ell}$, all variables, parameters and constraints of this model can be interpreted as in the static case. The objective function is the demand covered twice within r_1 time units minus the sum of penalties associated with ambulance relocations at time t.

Gendreau et al. (2001) solve DDSM^t by constructing a redeployment table in which the first column gives the list of all ambulances that could possibly be dispatched to the next call, and the second column gives the redeployment plan associated with the ambulance. Whenever an ambulance is dispatched, the associated relocation plan is implemented and the table is then recomputed from scratch. The authors have used the Gendreau et al. (1997) tabu search algorithm to compute the relocation strategies and they have also made use of a simple parallel computing strategy: each of 16 processors was assigned the computation of a series of rows of the redeployment table but no communication took place between the processors. The success of this approach rests on the capability of recomputing the entire table between successive calls. Simulations performed on real data from Montreal have shown that this was indeed possible in 95% of all cases. Out of all calls, 38% required at least one ambulance relocation and in only 0.05% of the situations were more than five ambulance relocations necessary. Whenever the table cannot be fully computed between two calls, it is deleted at the next call and no relocation takes place.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada under Grants 227837-00, OGP0039862, and 36662-01. This support is gratefully acknowledged. Most of the material in Section 5 is the result of discussions among the members of a project team designing and implementing scheduling technology for air taxi services, consisting of Mo Bazaraa, Marcos Coycoolea, Daniel Espinoza, Yongpei Guan, Renan Garcia, George Nemhauser, and Martin Savelsbergh from Georgia Tech and Alex Khmelnitsky and Eugene Taits from Jetson Systems.

References

Adams, T.F. (1997). *Police Field Operations*. Prentice Hall, Englewood Cliffs, NJ.

Alfa, A.S. (1986). Scheduling of vehicles for transportation of elderly. *Transportation Planning and Technology* 11, 203–212.

Attanasio, A., Cordeau, J.-F., Ghiani, G., Laporte, G. (2004). Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. *Parallel Computing* 30, 377–387.

Balas, E., Fischetti, M., Pulleyblank, W.R. (1995). The precedence-constrained asymmetric traveling salesman polytope. *Mathematical Programming* 68, 241–265.

Baldacci, R., Hadjiconstantinou, E., Mingozzi, A. (2003). An exact algorithm for the traveling salesman problem with deliveries and collections. *Networks* 42, 26–41.

Ball, M.O., Lin, L.F. (1993). A reliability model applied to emergency service vehicle location. *Operations Research* 41, 18–36.

Batta, R., Dolan, J.M., Krishnamurty, N.N. (1989). The maximal expected covering location problem: Revisited. *Transportation Science* 23, 277–287.

Bent, R., Van Hentenryck, P. (2004). Scenario-based planning for partially dynamic vehicle routing with stochastic customers. *Operations Research* 52, 573–586.

Benyahia, I., Potvin, J.-Y. (1998). Decision support for vehicle dispatching using genetic programming. *IEEE Transactions on Systems, Man and Cybernetics* 28, 306–314.

Bertsimas, D.J., van Ryzin, G.J. (1991). A stochastic and dynamic vehicle routing problem in the Euclidean plane. *Operations Research* 39, 601–615.

Bianco, L., Mingozzi, A., Ricciardelli, S., Spadoni, M. (1994). Exact and heuristic procedures for the traveling salesman problem with precedence constraints, based on dynamic programming. *INFOR* 32, 19–31.

Bodin, L.D., Sexton, T. (1986). The multi-vehicle subscriber dial-a-ride problem. *TIMS Studies in Management Science* 26, 73–86.

Borndörfer, R., Grötschel, M., Klostermeier, F., Küttner, C. (1999). Telebus Berlin: Vehicle scheduling in a dial-a-ride system. In: Wilson, N.H.M. (Ed.), *Computer-Aided Transit Scheduling*. *Lecture Notes in Economics and Mathematical Systems*, vol. 471. Springer-Verlag, pp. 391–422.

Brotcorne, L., Laporte, G., Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research* 147, 451–463.

Campbell, A., Savelsbergh, M.W.P. (2003). Incentive schemes for attended home delivery services. Technical Report TLI-03-04, Georgia Institute of Technology, The Logistics Institute.

Campbell, A., Savelsbergh, M.W.P. (2005). Decision support for consumer direct grocery initiatives. *Transportation Science* 39, 313–327.

Christofides, N. (1976). Worst case analysis of a new heuristic for the travelling salesman problem. Technical Report 388, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA.

Church, R.L., ReVelle, C.S. (1974). The maximal covering location problem. *Papers of the Regional Science Association* 32, 101–118.

Cordeau, J.-F. (2006). A branch-and-cut algorithm for the dial-a-ride problem. *Operations Research* 54, 573–586.

Cordeau, J.-F., Laporte, G. (2003a). The dial-a-ride problem (DARP): Variants, modeling issues and algorithms. *4OR – Quarterly Journal of the Belgian, French and Italian Operations Research Societies* 1, 89–101.

Cordeau, J.-F., Laporte, G. (2003b). A tabu search heuristic for the static multi-vehicle dial-a-ride problem. *Transportation Research B* 37, 579–594.

Cordeau, J.-F., Gendreau, M., Laporte, G. (1997). A tabu search heuristic for periodic and multi-depot vehicle routing problems. *Networks* 30, 105–119.

Coslovich, L., Pesenti, R., Ukovich, W. (2003). A two-phase insertion technique of unexpected customers for a dynamic dial-a-ride problem. Technical report, Università di Trieste, Italy.

Cullen, F.H., Jarvis, J.J., Ratliff, H.D. (1981). Set partitioning based heuristics for interactive routing. *Networks* 11, 125–143.

Daskin, M.S. (1983). A maximum expected location model: Formulation, properties and heuristic solution. *Transportation Science* 7, 48–70.

Daskin, M.S., Stern, E.H. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science* 15, 137–152.

Desaulniers, G., Desrosiers, J., Erdmann, A., Solomon, M.M., Soumis, F. (2002). VRP with pickup and delivery. In: Toth, P., Vigo, D. (Eds.), *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, PA, pp. 225–242.

Desrosiers, J., Dumas, Y., Soumis, F. (1986). A dynamic programming solution of the large-scale single-vehicle dial-a-ride problem with time windows. *American Journal of Mathematical and Management Sciences* 6, 301–325.

Desrosiers, J., Dumas, Y., Soumis, F., Taillefer, S., Villeneuve, D. (1991). An algorithm for mini-clustering in handicapped transport. Technical Report G-91-22, GERAD, HEC Montréal.

Dial, R.B. (1995). Autonomous dial-a-ride transit introductory overview. *Transportation Research C* 3, 261–275.

Dumas, Y., Desrosiers, J., Soumis, F. (1989). Large scale multi-vehicle dial-a-ride problems. Technical Report G-89-30, GERAD, HEC Montréal.

Dumas, Y., Soumis, F., Desrosiers, J. (1990). Optimizing the schedule for a fixed vehicle path with convex inconvenience costs. *Transportation Science* 24, 145–152.

Dumas, Y., Desrosiers, J., Soumis, F. (1991). The pickup and delivery problem with time windows. *European Journal of Operational Research* 54, 7–22.

Fischetti, M., Toth, P. (1989). An additive bounding procedure for combinatorial optimization problems. *Operations Research* 37, 319–328.

Fu, L. (2002). Scheduling dial-a-ride paratransit under time-varying, stochastic congestion. *Transportation Research B* 36, 485–506.

Fujiwara, O., Makjamroen, T., Gupta, K.K. (1987). Ambulance deployment analysis: A case study of Bangkok. *European Journal of Operational Research* 31, 9–18.

Garey, M.R., Johnson, D.S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, San Francisco.

Gendreau, M., Laporte, G., Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science* 5, 75–88.

Gendreau, M., Guertin, F., Potvin, J.-Y., Séguin, R. (1998). Neighborhood search heuristics for a dynamic vehicle dispatching problem with pick-ups and deliveries. Technical Report CRT-98-10, Centre de recherche sur les transports, Montreal.

Gendreau, M., Laporte, G., Vigo, D. (1999). Heuristics for the traveling salesman problem with pickup and delivery. *Computers & Operations Research* 26, 699–714.

Gendreau, M., Laporte, G., Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing* 27, 1641–1653.

Glover, F. (1996). Ejection chains, reference structures and alternating path methods for traveling salesman problems. *Discrete Applied Mathematics* 65, 223–253.

Goldberg, J., Dietrich, R., Chen, J.M., Mitwasi, M.G. (1990). Validating and applying a model for locating emergency medical services in Tucson, AZ. *European Journal of Operational Research* 49, 308–324.

Grötschel, M., Padberg, M.W. (1985). Polyhedral theory. In: Lawler, E.L., Lenstra, J.K., Rinnooy Kan, A.H.G., Shmoys, D.B. (Eds.), *The Traveling Salesman Problem*. Wiley, New York, pp. 251–305.

Healy, P., Moll, R. (1995). A new extension of local search applied to the dial-a-ride problem. *European Journal of Operational Research* 83, 83–104.

Hernández-Pérez, H., Salazar-González, J.-J. (2004). A branch-and-cut algorithm for a traveling salesman problem with pickup and delivery. *Discrete Applied Mathematics* 145, 126–139.

Hogan, K., ReVelle, C.S. (1986). Concepts and applications of backup coverage. *Management Science* 34, 1434–1444.

Holland, J.H. (1992). *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA.

Horn, M.E.T. (2002). Fleet scheduling and dispatching for demand-responsive passenger services. *Transportation Research C* 10, 35–63.

Hunsaker, B., Savelsbergh, M. (2002). Efficient testing for dial-a-ride problems. *Operations Research Letters* 30, 169–173.

Ioachim, I., Desrosiers, J., Dumas, Y., Solomon, M.M. (1995). A request clustering algorithm for door-to-door handicapped transportation. *Transportation Science* 29, 63–78.

Jaw, J., Odoni, A.R., Psaraftis, H.N., Wilson, N.H.M. (1986). A heuristic algorithm for the multi-vehicle advance-request dial-a-ride problem with time windows. *Transportation Research B* 20, 243–257.

Kalantari, B., Hill, A.V., Arora, S.R. (1985). An algorithm for the traveling salesman problem with pickup and delivery customers. *European Journal of Operational Research* 22, 377–386.

Koza, J.R. (1992). *Genetic Programming*. MIT Press, Cambridge, MA.
Larson, R.C. (1972). *Urban Police Patrol Analysis*. MIT Press, Cambridge, MA.
Larson, R.C. (1974). A hypercube qeueing model for facility location and redistricting in urban emergency services. *Computers & Operations Research* 1, 67–75.
Lau, H.C., Liang, Z. (2002). Pickup and delivery with time windows: Algorithms and test case generation. *International Journal on Artificial Intelligence Tools* 11, 455–472.
Leclerc, F., Potvin, J.-Y. (1997). Genetic algorithms for vehicle dispatching. *International Transactions in Operational Research* 4, 391–400.
Lin, S. (1965). Computer solutions of the travelling salesman problem. *Bell System Technical Journal* 44, 2245–2269.
Lin, S., Kernighan, B.W. (1973). An effective heuristic algorithm for the traveling-salesman problem. *Operations Research* 21, 498–516.
Little, J.D.C., Murty, K.G., Sweeney, D.W., Karel, C. (1963). An algorithm for the traveling salesman problem. *Operations Research* 11, 972–989.
Madsen, O.B.G., Ravn, H.F., Rygaard, J.M. (1995). A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of Operations Research* 60, 193–208.
Mandell, M.B. (1998). Covering models for two-tiered emergency medical services systems. *Location Science* 6, 355–368.
Marianov, V., ReVelle, C.S. (1994). The queueing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Sciences* 28, 167–178.
Marianov, V., ReVelle, C.S. (1995). Siting emergency services. In: Drezner, Z. (Ed.), *Facility Location*. Springer-Verlag, New York, pp. 199–223.
Mitrović-Minić, S., Laporte, G. (2004). Waiting strategies for the dynamic pickup and delivery problem with time windows. *Transportation Research B* 38, 635–655.
Mitrović-Minić, S., Krishnamurti, R., Laporte, G. (2004). Double-horizon based heuristics for the dynamic pickup and delivery problem with time windows. *Transportation Research B* 38, 669–685.
Nanry, W.P., Barnes, J.W. (2000). Solving the pickup and delivery problem with time windows using reactive tabu search. *Transportation Research B* 34, 107–121.
Psaraftis, H.N. (1980). A dynamic programming approach to the single-vehicle, many-to-many immediate request dial-a-ride problem. *Transportation Science* 14, 130–154.
Psaraftis, H.N. (1983a). Analysis of an $O(N^2)$ heuristic for the single vehicle many-to-many Euclidean dial-a-ride problem. *Transportation Research B* 17, 133–145.
Psaraftis, H.N. (1983b). An exact algorithm for the single-vehicle many-to-many dial-a-ride problem with time windows. *Transportation Science* 17, 351–357.
Psaraftis, H.N. (1983c). k-interchange procedures for local search in a precedence-constrained routing problem. *European Journal of Operational Research* 13, 391–402.
Psaraftis, H.N. (1988). Dynamic vehicle routing problems. In: Golden, B.L., Assad, A.A. (Eds.), *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, pp. 223–248.
Rego, C., Roucairol, C. (1996). A parallel tabu search algorithm using ejection chains for the vehicle routing problem. In: Osman, I.H., Kelly, J.P. (Eds.), *Meta-Heuristics: Theory & Applications*. Kluwer Academic, Boston, pp. 661–675.
Renaud, J., Boctor, F.F., Laporte, G. (1996). A fast composite heuristic for the symmetric traveling salesman problem. *INFORMS Journal on Computing* 8, 134–143.
Renaud, J., Boctor, F.F., Ouenniche, J. (2000). A heuristic for the pickup and delivery traveling salesman problem. *Computers & Operations Research* 27, 905–916.
Renaud, J., Boctor, F.F., Laporte, G. (2002). Perturbation heuristics for the pickup and delivery traveling salesman problem. *Computers & Operations Research* 29, 1129–1141.
Repede, J.F., Bernardo, J.J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research* 75, 567–581.
ReVelle, C.S., Hogan, K. (1989). The maximum availability location problem. *Transportation Science* 23, 192–200.

Rochat, Y., Taillard, É.D. (1995). Probabilistic diversification and intensification in local search for vehicle routing. *Journal of Heuristics* 1, 147–167.

Ropke, S., Pisinger, D. (2004). An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows. Technical Report 2004-13, DIKU, University of Copenhagen.

Ruland, K.S. (1995). Polyhedral solution to the pickup and delivery problem. PhD thesis, Sever Institute of Technology, Washington University.

Ruland, K.S., Rodin, E.Y. (1997). The pickup and delivery problem: Faces and branch-and-cut algorithm. *Computers and Mathematics with Applications* 33, 1–13.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), *Parallel Distributed Processing, vol. 1*. MIT Press, Cambridge, MA, pp. 318–364.

Savelsbergh, M.W.P. (1985). Local search in routing problems with time windows. *Annals of Operations Research* 4, 285–305.

Savelsbergh, M.W.P. (1990). An efficient implementation of local search algorithms for constrained routing problems. *European Journal of Operational Research* 47, 75–85.

Savelsbergh, M.W.P. (1992). The vehicle routing problem with time windows: Minimizing route duration. *ORSA Journal on Computing* 4, 146–154.

Savelsbergh, M.W.P., Sol, M. (1995). The general pickup and delivery problem. *Transportation Science* 29, 17–29.

Savelsbergh, M.W.P., Sol, M. (1998). DRIVE: Dynamic routing of independent vehicles. *Operations Research* 46, 474–490.

Schilling, D.A., Elzinga, D.J., Cohon, J., Church, R.L., ReVelle, C.S. (1979). The TEAM FLEET models for simultaneous facility and equipment siting. *Transportation Science* 13, 163–175.

Sexton, T., Bodin, L.D. (1985a). Optimizing single vehicle many-to-many operations with desired delivery times: I. Scheduling. *Transportation Science* 19, 378–410.

Sexton, T., Bodin, L.D. (1985b). Optimizing single vehicle many-to-many operations with desired delivery times: II. Routing. *Transportation Science* 19, 411–435.

Shen, Y., Potvin, J.-Y., Rousseau, J.-M., Roy, S. (1995). A computer assistant for vehicle dispatching with learning capabilities. *Annals of Operations Research* 61, 189–211.

Stein, D.M. (1978). Scheduling dial-a-ride transportation systems. *Transportation Science* 12, 232–249.

Swersey, A.J. (1994). The deployment of police, fire, and emergency medical units. In: Pollock, S.M., Rothkopf, M.H., Barnett, A. (Eds.), *Operations Research and the Public Sector*. *Handbooks in Operations Research and Management Science*, vol. 6. North-Holland, pp. 151–200.

Swihart, M.R., Papastravou, J.D. (1999). A stochastic and dynamic model for the single-vehicle pick-up and delivery problem. *European Journal of Operational Research* 114, 447–464.

Teodorovic, D., Radivojevic, G. (2000). A fuzzy logic approach to dynamic dial-a-ride problem. *Fuzzy Sets and Systems* 116, 23–33.

Toregas, C.R., Swain, R., ReVelle, C.S., Bergman, L. (1971). The location of emergency service facilities. *Operations Research* 19, 1363–1373.

Toth, P., Vigo, D. (1996). Fast local search algorithms for the handicapped persons transportation problem. In: Osman, I.H., Kelly, J.P. (Eds.), *Meta-Heuristics: Theory & Applications*. Kluwer Academic, Boston, pp. 677–690.

Toth, P., Vigo, D. (1997). Heuristic algorithms for the handicapped persons transportation problem. *Transportation Science* 31, 60–71.

Van der Bruggen, L.J.J., Lenstra, J.K., Schuur, P.C. (1993). Variable-depth search for the single-vehicle pickup and delivery problem with time windows. *Transportation Science* 27, 298–311.

Walker, W.E., Chaiken, J.M., Ignall, E.J. (1979). *Fire Department Deployment Analysis*. North-Holland, New York.

Xu, H., Chen, Z.-L., Rajagopal, S., Arunapuram, S. (2003). Solving a practical pickup and delivery problem. *Transportation Science* 37, 347–364.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14008-6

Chapter 8

Intermodal Transportation

Teodor Gabriel Crainic

Département management et technologie, École des Sciences de la Gestion, Université du Québec à Montréal, and Centre de recherche sur les transports – CIRRELT, Montréal, Canada
E-mail: theo@crt.umontreal.ca

Kap Hwan Kim

Department of Industrial Engineering, Pusan National University, Korea
E-mail: kapkim@pusan.ac.kr

Keywords: Intermodal transportation, freight transportation, operations research

1 Introduction

Intermodal transportation may be defined as the transportation of a person or a load from its origin to its destination by a sequence of at least two transportation modes, the transfer from one mode to the next being performed at an intermodal terminal. The concept is very general and thus, it means many things to many people: transportation of containerized cargo by a combination of truck, rail, and ocean shipping, dedicated rail services to move massive quantities of containers and trailers over long distances, main transportation mode for the international movement of goods, central piece in defining transportation policy for the European Community, trips undertaken by a combination of private (e.g., car) and public (e.g., light rail) transport, and so on. One must therefore start with a few definitions to set the terminology and limit the scope of this chapter. First, although both people and freight transportation can be examined from an intermodal perspective, we limit the scope of this chapter to freight.

In one of its most widely accepted meanings, intermodal freight transportations refers to a *multimodal* chain of *container*-transportation services. This chain usually links the initial shipper to the final consignee of the container (so-called *door-to-door* service) and takes place over long distances. Transportation is often provided by several carriers. In a classical example of an intercontinental intermodal chain, loaded containers leave a shipper's facility by truck either directly to port or to a rail yard from where a train will deliver them to port. A ship will move the containers from this initial port to a port on the other continent, from where they will be delivered to the final destination by a single or a combination of "land" transportation means: truck, rail, coastal or river

navigation. Several intermodal terminals are part of this chain: the initial and final seaport container terminals, where containers are transferred between the ocean navigation and land transportation modes, as well as in-land terminals (rail yards, river ports, etc.) providing transfer facilities between the land modes.

Container transportation is a major component of intermodal transportation and international commerce and this importance is reflected in this chapter. Intermodal transportation is not only about containers and intercontinental exchanges, however. On the one hand, a significant part of international trade that is moved in containers does not involve ocean navigation, land transportation means providing the intermodal chain. On the other hand, other types of cargo may be moved by a chain of transportation means and require intermodal transfer facilities, as illustrated by the definition European Conference of Ministers of Transport (2001) gives for intermodal transportation: "movement of goods in one and the same loading unit or vehicle, which uses successively two or more modes of transport without handling the goods themselves in changing modes". This last definition is still too restrictive, however. Thus, for example, the transportation of express and regular mail on a regional or national scale is strongly intermodal, using various combinations of road, rail, and air transportation modes, and yet freight is handled (sorted and grouped) in terminals. More generally, the transportation of less-than-vehicle-capacity loads by nondedicated services is intermodal, since it involves pickup (at origin) and delivery (at destination) operations, usually performed by trucks, at least one long-haul transportation movement by road, rail, river, or air, as well as transfer activities between these modes in dedicated terminals.

Almost all types of freight carriers and terminal operator may, thus, be involved in intermodal transportation, either by providing service for part of the transportation chain or by operating an intermodal transportation system (network). We limit the scope of the chapter to the latter with a particular focus on container-based systems, including container terminals in seaports. The national/regional planning perspective, which considers the flow of multiple products on multimodal networks, is also addressed.

Compared to several other application areas, Operations Research models and methods for intermodal freight transportation is still a very young domain. In many cases, there are not, yet, widely accepted models and methodologies. Work is indeed proceeding as this chapter is being written. Therefore, the goal of the chapter is to be informative and provide a starting point for future research. The chapter overviews the evolution of the intermodal transportation field and presents methodological developments proposed to address a number of important operations and planning issues: system and service design for intermodal transportation networks, container fleet management, container terminal operations and scheduling, national planning. We focus on models. Algorithmic developments are indicated but not examined in any depth.

To structure the presentation, we follow the somewhat classical approach of examining issues, models, and methods according to whether they belong to

the *strategic*, *tactic*, or *operational* level of planning and management of operations. The chapter is therefore organized as follows. Section 2 briefly describes freight transportation and the main actors and issues discussed in the chapter. Section 3 is dedicated to system and service network design issues for carriers. Strategic design issues for seaport container terminals are also discussed in this section. Operational-planning issues are addressed in Section 4 with particular emphasis on empty container repositioning. Section 5 is dedicated to tactical and operational planning of operations in container terminals, while Section 6 examines national planning models. Section 7 concludes the chapter with an enumeration of a number of trends and technological developments that may influence intermodal transportation in the future, as well as of promising research directions.

2 Freight transportation systems

Demand for freight transportation derives from the interplay between *producers* and *consumers* and the significant distances that often separate them. Producers of goods require transportation services to move raw materials and intermediate products, and to distribute final goods in order to meet customer demands. *Shippers*, which may be the producers of goods or some intermediary firm (e.g., *brokers*), thus generate the *demand* for transportation. *Carriers* answer this demand by *supplying* transportation services. Railways, ocean shipping lines, trucking companies, and postal services are examples of carriers. Considering the type of services they provide, seaports, intermodal platforms, and other such facilities may be described as carriers as well. Governments contribute the infrastructure: roads and highways, as well as significant portions of ports, internal navigation, and rail facilities. Governments also regulate (e.g., dangerous and toxic goods transportation) and tax the industry.

We do not intend to make a detailed presentation of freight transportation. Our goal is rather to describe the basic operation and planning issues for long-haul carriers and terminals that are involved in intermodal transportation, with a particular focus on container-based systems, including container terminals in seaports. The first subsection gives a few statistics and trends relative to container-based transportation. The second and the third are dedicated to long-haul carriers and seaport container terminals, respectively.

Reviews on these issues may be found in Assad (1980), Cordeau et al. (1998), and Crainic (1988) for rail transportation, Delorme et al. (1988) and Powell (1988) for motor carrier transportation, Christiansen et al. (2004, 2007) for maritime transportation, and Günther and Kim (2005) and Steenken et al. (2004) for container port terminals. The reviews proposed by Crainic (2000, 2003), Crainic and Laporte (1997), Daganzo (2005), Dejax and Crainic (1987), Powell (2003), Powell and Topaloglu (2003, 2005), Powell et al. (1995, 2007) are more general in scope, addressing issues relevant for more than one transportation mode or problem. Issues and methodologies related to the pickup

and delivery of loads, generally known as vehicle routing and scheduling problems, are not included in this chapter. Interested readers may refer to Golden and Assad (1988), Ball et al. (1995), Dror (2000), Toth and Vigo (2002), and Cordeau et al. (2007).

2.1 Container intermodal transportation

Container-related transportation activities have grown remarkably over the last 10 years and the trend does not show any sign of slowing down as illustrated by the annual world container traffic figures, in millions of TEUs (20 feet equivalent container units), displayed in Table 1 (2006 figures are estimated; Koh and Kim, 2001; ISL, 2006). The initial impulse to container-based transportation came from the safety it offered regarding loss and damage. Advantages in terms of reduced cargo handling and standardization of transportation and transfer equipment translate in cost economies and efficient, world-wide door-to-door intermodal service and fuel the growth of the industry. Containerized intermodal transportation supports a significant part of the international movement of goods.

The performance of container-based transportation in international trade has had some remarkable consequences. Ports and container terminals have been built or profoundly modified to accommodate container ships and efficiently perform the loading, unloading, and transfer operations. Container terminal equipment and operating procedures are continuously enhanced to improve productivity and compete, in terms of cost and time, with the other ports to attract ocean-shipping lines. Models for planning container-terminal operations are presented in Section 5. Notice that the competitive position of a container port is also dependent on the capacity and efficiency of the land transportation system. The models described in Section 6 may be used for the integrated analysis and planning of port and land transportation systems.

Table 1.
World container traffic

Year	Container traffic (Millions TEU)	Growth rate (%)
1993	113.2	12.5
1995	137.2	9.8
1997	153.5	4.2
1999	203.2	10
2000	225.3	10.9
2001	231.6	2.8
2002	240.6	3.9
2003	254.6	5.8
2004	280.0	10.6
2005	310.5	10.9

With respect to ocean navigation, efficiency reasons have led to the construction of very large container-carrying ships for intercontinental movements; new container ship capacities are of the order of 8000 to 10,000 TEUs. To operate efficiently, such ships must not stop frequently. Moreover, they are too large for the vast majority of ports. Consequently, a new link has been added to the intermodal chain: super-ships stop at a small number of major seaports and containers are transferred to smaller ships for distribution to various smaller ports. Notice that these ships cannot navigate through the Panama Canal. This results in a modification of sea and land shipping routes. The impact of these modifications on particular transportation systems and facilities may be analyzed by using the models of Section 6.

The impact of the growth in containerized trade has been significant on land transportation systems as well. Specialized transportation services have been created, such as the North-American "land-bridges" that provide container transportation by long, double-stack trains operated by independent subsidiaries of the rail companies between the East and the West coasts and between these ports and the industrial core of the continent. Dedicated rail services are also being created in Europe to provide shuttle services for container transportation. Section 3.2.3 examines some of the issues associated to planning and operating such dedicated services.

2.2 Customized, consolidation, and intermodal transportation

In an intermodal chain, one may encounter *consolidation* transportation systems, where one vehicle or convoy serves to move freight for different customers with possibly different initial origins and final destinations, and *customized* transportation carriers that provide dedicated service to each particular customer.

Truckload trucking offers a typical example of customized transportation. When the customer calls, the dispatcher assigns to the task a truck and driver (or driving team for very long movements). The truck moves to the customer-designated location, is loaded, and then moves to the specified destination where it is unloaded. The driver then calls the dispatcher to give its position and request a new assignment. The dispatcher may indicate a new load, ask the driver to move empty to a new location where demand should appear in the near future, or have the driver wait and call later.

The truckload carrier thus operates in a highly dynamic environment, where little is known with certainty regarding future demands, waiting delays at customer locations, precise positions of loaded and empty vehicles at later moments in time, and so on. Moreover, the time available to the dispatcher to decide on the next assignment in response to a customer or driver request, is generally very short (most such decisions are performed in real time). Service is tailored for each customer and the timely assignment of vehicles to profitable demands is very important. The development of efficient *resource management and allocation* strategies is therefore at the heart of the management process.

These strategies attempt to maximize the volume of demand satisfied (loads moved) and the associated profits, while making the best use of the available resources: drivers, tractor and trailer fleets, etc. We do not address these issues in this chapter. The interested reader may consult Powell (2003) or Powell et al. (2007), for example. Notice, however, that most methodology targeting resource management and allocation issues for customized carriers may also be adapted to address operational-level fleet management issues for consolidation carriers (Section 4).

Ocean navigation services provided by for-hire ships share most of these dynamic and stochastic characteristics. Variations in travel times are in fact larger for sea than for road trips. On the other hand, travel and loading/unloading times are usually much longer, which increases the time available to decide on the next assignment. Railways also provide customized services by dedicating unit trains to individual shippers. Such decisions respond in most cases to long-term contracts, which implies that demand is known deterministically. This contrasts to the stochastic demand that characterizes truckload trucking and for-hire ocean shipping. The setup of such unit trains is therefore part of the regular planning process of operations as described later on in this section.

Customized transportation is not always the appropriate answer to shipper needs. The relations and trade-offs between volume and frequency of shipping, on the one hand, and the cost, frequency, and delivery time of transportation, on the other hand, often dictates the use of *consolidation* transportation services. In passenger transportation, this choice is equivalent, for example, between taking a taxi or using the services of a regular public transport line. Freight consolidation transportation is performed by Less-Than-Truckload (LTL) motor carriers, railways, ocean shipping lines, regular and express postal services, etc. Freight transportation in some countries where a central authority more or less controls a large part of the transportation system also belongs to this category.

Consolidation transportation carriers and fundamentally all intermodal transportation systems are organized as so-called *hub-and-spoke* networks, illustrated in Figure 1. In such systems, service is offered between a certain number of origin–destination points (the local/regional terminals), represented by nodes 1–9 in Figure 1. This number is significantly larger than the number of direct, origin to destination services operated by the carrier. Consequently, and to take advantage of economies of scale, low-volume demands are moved first to an intermediate point – a consolidation terminal or a hub (nodes A, B, and C in Figure 1) – such as an airport, seaport container terminal, rail yard, or intermodal platform. At a hub, traffic is consolidated into larger flows that are routed to other hubs by high-frequency, high-capacity services (thick full and dashed lines in Figure 1). As illustrated in the figure, more than one service, of possibly different modes, may be operated between hubs. Lower frequency services, often operating smaller vehicles, are used between hubs and origin–destination terminals (the regular lines in the figure). When the level of demand justifies it, high-frequency, high-capacity services may be run between

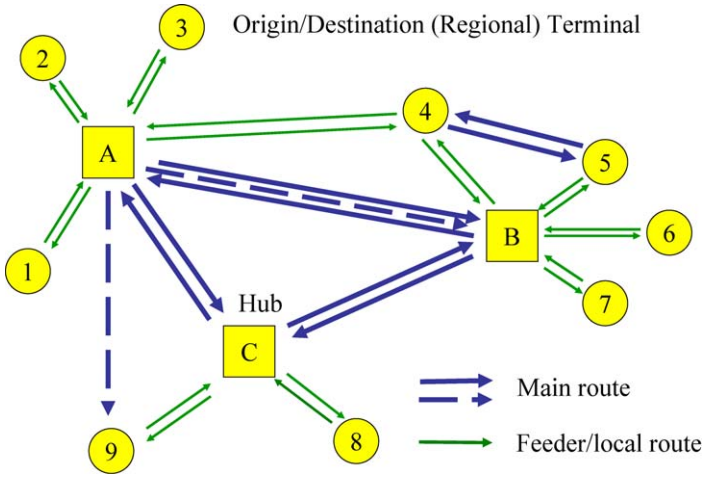


Fig. 1. Network with consolidation terminals/hubs.

a hub and a regional terminal (link between nodes A and 9) or between two regional terminals (links between nodes 4 and 5). Notice that cargo is brought to and distributed from origin–destination terminals by vehicles performing either pickup and distribution routes or customized services. LTL motor carriers and postal services are examples of the first case, rail and ocean shipping of containers (barges or trucks performing the local transportation activities) of the second. Yet, because the planning of the long-haul activities of consolidation carriers does not generally include these activities, we do not elaborate further on the topic.

A hub-and-spoke organization allows a much higher frequency of service between all origin–destination pairs in the network and a more efficient utilization of resources. The drawback of this type of organization is increased delays due to longer routes and the time spent in terminals. This explains partly why there are very few "pure" hub-and-spoke systems in operation; direct transportation is usually operated for high-demand or high-priority origin–destination pairs. A second strategy is to separate high- and low-priority traffic and to dedicate different services and, eventually, infrastructure (e.g., terminals) to each. An example of this trend is the creation of intermodal subdivisions by North-American railways to ensure efficient movements of containers among the major ports and industrial centers of the continent.

To further mitigate the drawbacks of hub-based operations, consolidation carriers engage into rather sophisticate planning activities (Section 3). The carrier operates a series of *services*, each characterized by its own route, stops, frequency, vehicle and convoy type, capacity, speed (travel time), and so on. Internally, services are often collected in an *operational plan* (also referred to as *load* or *transportation* plan), generally accompanied by a *schedule* that indicates departure and arrival times at the terminals of the route. The schedule

is partially (e.g., latest delivery of cargo at the origin terminal for on-time delivery at destination) or totally available to customers. The aim of the load plan is to ensure that the proposed services are performed as stated (or as closely as possible), while operating in a rational, efficient, and profitable way. It also indicates how demand is moved though the system using its terminals and transportation services.

Freight demand is defined by its specific origin, destination, and commodity-related physical characteristics (e.g., weight and volume), as well as particular service requirements in terms of delivery conditions, type of vehicle, and so on. A profit or cost is also usually associated to a given demand. Once delivered at the carrier's terminal, the cargo of several customers is sorted, grouped, and loaded into the same vehicle or convoy. It is then moved either directly, when such a service exists, or through a series of services with intermediary operations of transfer and consolidation. More than one consolidation-transfer operation may occur during a trip. In the case of LTL motor carriers, and often for postal services, it is individual loads that are consolidated and loaded into vehicles, trucks, planes, or rail cars. Containerized cargo is not handled before reaching its destination and, thus, consolidation operations involve only the containers, which are loaded into ships, airplanes, or rail cars. This constitutes a first type of consolidation operation: "small" loads into vehicles, containers into ships, etc. In some cases, vehicles are further sorted and consolidated into convoys. The most widely-spread and well-known case is that of railway transportation where, first, cars are grouped into blocks (i.e., the cars in a block travel together, without any re-consolidation operation, until some common, intermediary or final, terminal) and, then, blocks are put together to make up trains. Similar make-up operations, albeit on a smaller scale, also occur for barge trains and multitrailer assemblies.

Terminals are clearly an important component of consolidation and intermodal transportation systems, and their efficiency is vital to the performance of the entire transport chain. The following subsection briefly describes main terminal operations and planning issues with a particular focus on container port terminals.

2.3 Consolidation and intermodal terminals

Terminals come in several designs and sizes and may be specialized for particular transportation modes and the handling of specific products, or may offer a complete set of services. Major operations performed in consolidation terminals include vehicle loading and unloading, cargo and vehicle sorting and consolidation, convoy make up and break down, and vehicle transfer between services.

When containerized traffic is concerned, only the containers are handled, not the cargo they contain. Thus, once loaded on a truck, rail car, barge, or ocean vessel, containers will generally follow the movements and consolidation activities of the respective vehicle, until reaching either the final destination or

an intermodal terminal. There are very few such activities for motor carriers and barges. We may mention the transfer of a barge or a trailer from a barge or road train to another. No particular planning model is built for such activities; rather, these are included in the scope of the models used to build the transportation plan of the carrier (Section 3.2).

More complex consolidation operations are performed within railway systems: sorting and consolidation of rail cars into blocks and trains. A rich literature exists on models targeting these issues and is reviewed in the references indicated at the beginning of the section. In most cases, however, there are no differences between intermodal and regular rail traffic with respect to blocking and train make up planning and operations, even when particular terminals are dedicated to handling intermodal traffic. Research is under way to develop the "next generation" intermodal rail services and terminals, but the efforts dedicated to the associated planning issues are still limited (e.g., Bostel and Dejax, 1998; Macharis and Bontekoning, 2004).

The intermodal transfer of containers between truck and rail, taking place at rail terminals, is specific to intermodal transportation. Containers thus arrive at the rail terminal by truck and are either directly transferred to a rail car or, more frequently, are stacked in a waiting area. Then, containers are picked up from the waiting area and loaded unto rail cars that will be grouped into blocks and trains. The reverse operations take place when containers arrive by train to the terminal and are to be transfered to trucks for their next transport leg. We did not find any paper dedicated to the planning of these operations. The issues are very similar to those arising in container port terminals, however, which are probably the most well-known intermodal transfer facilities. Significant research has been dedicated to the design (Section 3.3) and particularly to the operations (Section 5) of container port terminals. In the following, we briefly describe these operations and planning issues.

The main function of a container port terminal is to provide transfer facilities for containers between sea vessels and land transportation modes, trucks and rail in particular. It is a highly complex system that involves numerous pieces of equipment, operations, and container handling steps (Steenken et al., 2004). The assignment of resources to tasks and the scheduling of these tasks are thus among the major container port terminal planning issues. Three main areas make up a container terminal. The *sea-side* area encompasses the quays where ships berth and the quay cranes that provide the loading and unloading of containers into and from ships. The *land-side* area provides the interface with the land transportation system (the so-called hinterland of the port) and encompasses the truck and train receiving gates, the areas where rail cars are loaded and unloaded, and the associated equipment. Trucks are generally loaded and unloaded directly in the *yard area*. This third area is dedicated for the most part to stacking loaded and empty containers for import and export (in some terminals, facilities are also provided for the loading and loading of containers). Various types of yard cranes are associated with this area. So-called *transporters*, primarily yard trucks or automated vehicles, move con-

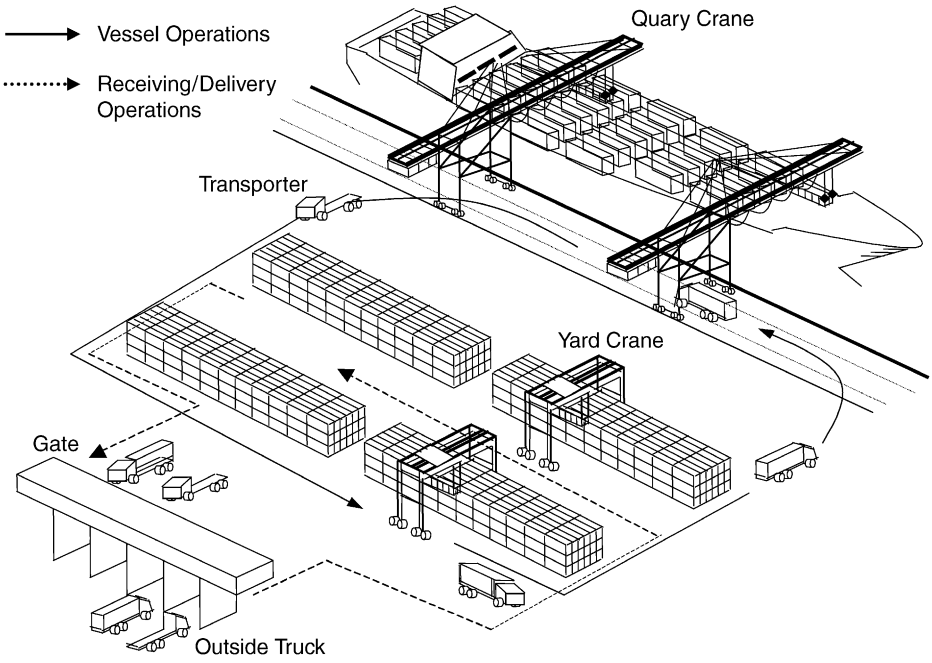


Fig. 2. Example of a container terminal with an indirect transfer system (Park, 2003).

tainers between the three areas. Figure 2 illustrates part of a container port terminal. One ship and three quay cranes are displayed in the sea-side area, while only trucking is shown in the land-side area. Twelve container stacks are displayed in the yard area, as well as one type of yard crane used to transfer containers between yard transporters and outside trucks and stacks, as well as to change the position of containers in the yard as required.

Three main types of handling operations are performed in a container terminal:

(1) ship operations associated with berthing, loading, and unloading container ships,

(2) receiving/delivery operations for outside trucks and trains, and

(3) container handling and storage operations in the yard.

When a ship arrives at the container port terminal, it is assigned a berth and a number of quay cranes. Berth space is a very important resource in a container terminal (construction costs to increase capacity are very high, even when space for growth exists) and *berth scheduling* determines the berthing time and position of a container ship at a given quay (Section 5.1). *Quay-crane allocation* is the process of determining the vessel that each quay crane will serve and the associate service time (Section 5.2). *Stowage sequencing* determines the sequence of unloading and loading containers, as well as the precise position each container being loaded into the ship is to be placed *in* (Section 5.3). Dur-

ing the unloading operation, a quay crane transfers a container from a ship to a transporter. Then, the transporter delivers the import (unloading) container to a yard crane that picks it up and stacks it into a given position in the yard. This sequence of operations is called *indirect transfer*. Some terminals use a *direct transfer* system where the equipment used to move containers between the quay and the yard will also stack them. For export (loading) operation, the process is carried out in the opposite direction.

On the land-side, the *receiving* and *delivery* operations provide the interface between the container terminal activities and the external movements. A receiving operation starts when containers arrive at the gate of the terminal carried by one or several outside trucks or a train. Containers are inspected at the gate to check for damages (to the container not its content) and whether all documents are in order. Also at the gate, information regarding where the container is to be stored is provided to the truck driver. When the outside truck arrives at the indicated transfer point, a yard crane lifts a container from the truck and stacks it according to the plan. When containers arrive by rail, the rail cars are brought to the rail area where containers and documents are examined. Containers are then transferred by a gantry crane to a transporter, which delivers them to the yard and stacks them. In the case of a delivery operation, the yard equipment delivers a container onto an outside truck, which leaves the port, or onto a transporter which delivers the container to the rail area and loads it onto the designated car.

The sea- and land-side operations interact with the yard container handling and storage operation through the information on where the containers are or must be stacked within the yard. How containers are stored in the yard is one of the important factors that affect the turn-around time of ships and land vehicles. The *space-allocation* problem is concerned with determining storage locations for containers either individually or as a group. Yard storage space is pre-assigned to containers of each ship arriving in the near future to maximize the productivity of the loading and unloading operations (Section 5.4).

A container yard consists of blocks of containers, which are separated by aisles for transporters as shown in Figure 2. A block consists of 25–35 yard bays, and a yard bay has 6–10 stacks of containers. Container handling and storage operations include the management and handling of containers while they are in storage in the yard and thus occur between the receiving and delivery operations and the ship operations. Container-handling equipment performs the placement of containers into storage and their retrieval when needed. Yard cranes move along blocks of containers to yard bays to perform these operations. Planning these operations is part of the *equipment-assignment* process, which allocates tasks to container-handling equipment. Based on the quay-crane schedule, one or two yard cranes are assigned to each quay crane for loading and unloading. The remaining yard cranes are allocated to receiving and delivery operations. Terminal operators aim to assign and operate yard cranes in such a way that inefficient moves and interferences among yard cranes are minimized (Section 5.5).

3 System and service network design

This section is dedicated to models aimed at strategic and tactical planning issues for freight carriers and seaport container terminals. At the strategic level, we focus on decisions that concern the design of the physical infrastructure network:

where to locate terminals (e.g., consolidation terminals, rail yards, intermodal platforms, and so on)?
what type and quantity of equipment (e.g., cranes) to install at each facility?
what type of lines or capacity to add?
what lines or facilities to abandon?
which customer zones to serve directly and how to serve the others?
and so on.

We group these issues under the label *system design*. The term *service network design* covers tactical planning issues for consolidation transportation firms:

on what routes to provide service?
what type of service (mode) to use?
how often to offer service on each route and according to what schedule?
how to route the loads through the physical and service networks?
how to distribute the work among the terminals of the system?

The output of the process is a transportation (load) plan. The goals are customer satisfaction and cost-efficient utilization of resources (assets) leading to profits.

Most freight carriers face system and service network design issues, irrespective of their involvement in intermodal transportation activities. It is not our intention, however, to present a comprehensive and detailed treatment of the subject. We focus rather on the main issues and modeling efforts that bear directly on intermodal transportation.

3.1 Models for system design

The literature on system design models for freight transportation is not rich. Such issues are often addressed by evaluating alternatives using network models for tactical or operational planning of transportation activities. When formal models are proposed, they generally take the form of discrete location formulations to address issues related to the location of consolidation or hub terminals and the routing of demand from its origin to its destination terminals. The routing of flows determines the direct connections (physical infrastructure or service links) between origins, destinations, and consolidation terminals. When these connections must be explicitly decided (e.g., the allocation of "local" terminals to major classification facilities), a combined

location-network-design formulation is often used. All formulations aim to capture the potential economies of scale associated with the consolidation of freight.

An extensive literature exists on location and network design models and solution methods: Mirchandani and Francis (1990), Daskin (1995), Drezner (1995), Labbé et al. (1995), Labbé and Louveaux (1997), Crainic and Laporte (1997), Drezner and Hamacher (2002), and Daskin and Owen (2003) review location issues and literature, while Magnanti and Wong (1984), Minoux (1989), Nemhauser and Wolsey (1993), Salkin and Mathur (1989), Ahuja et al. (1993), Balakrishnan et al. (1997), and Crainic (2000) survey the network design field.

In the following, we focus on the contributions related to intermodal transportation, the location of consolidation facilities in particular. All the models presented are static and deterministic location formulations. They assume that all problem components, particularly the demand as well as the cost and profit structure, will not vary during the planning period for which their evaluation is performed, usually from six to twelve months. Moreover, problems where the design decisions have long-term effects (e.g., location of hubs) also assume that variances will be negligible for the foreseeable future. These observations emphasize the need for research into time-dependent stochastic models for system design problems. Most problems presented in this subsection are NP-hard and thus heuristics are the solution method of choice in almost all cases.

3.1.1 Location with balancing requirements

The multicommodity location problem with balancing requirements was first introduced by Crainic et al. (1989) in the context of the management of a heterogeneous fleet of containers by an international maritime shipping company. The land operations of the carrier proceed as follows. Once a ship arrives at port, the company has to deliver loaded containers, which may come in several types and sizes, to designated in-land destinations. Following their unloading by the importing customer, empty containers are moved to a depot. From there, empty containers may be delivered to customers which request them for subsequent shipping of their own products. In addition, empty containers are often moved, or *repositioned*, to other depots. These interdepot movements are a consequence of regional imbalances in empty container availabilities and needs throughout the network: some areas lack containers of certain types, while others have surpluses. These balancing movements of empty containers among depots differentiate this problem from classical location–allocation applications. Interdepot movements are performed by high-density transport (rail, typically) and their unit transportation cost is lower than for the other types of movements. The general problem is therefore to locate depots and allocate customers to depots (for each type of container and direction of movement) in order to collect the supply of empty containers available at customers' sites and to satisfy the customer requests for empty containers, while minimiz-

ing the total operating costs: the costs of opening and operating the depots, and the costs generated by customer-depot and interdepot movements.

The formulation proposed by Crainic et al. (1989) is based on a directed network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ is the set of nodes or vertices and $\mathcal{A}$ is the set of arcs or links. The set of nodes is partitioned into three subsets: $\mathcal{O}$, the set of origin nodes (supply customers); $\mathcal{D}$, the set of destination nodes (demand customers); and $\mathcal{H}$, the set of hubs or consolidation nodes (depots). The sets of customers adjacent to each depot $j \in \mathcal{H}$ are identified as $\mathcal{O}(j) = \{i \in \mathcal{O}\colon (i, j) \in A\}$ and $\mathcal{D}(j) = \{i \in \mathcal{D}\colon (j, i) \in A\}$. It is assumed that there exists at least one origin or destination adjacent to each depot j ($\mathcal{O}(j) \cup \mathcal{D}(j) \neq \emptyset$). The sets of depots adjacent to each node $i \in N$ in both directions are defined as $\mathcal{H}^+(i) = \{j \in \mathcal{H}\colon (i, j) \in A\}$ and $\mathcal{H}^-(i) = \{j \in \mathcal{H}\colon (j, i) \in A\}$. The problem description does not allow direct customer-to-customer movements of containers. Hence, the set of arcs $\mathcal{A}$ may be partitioned into three subsets: customer-to-depot arcs, $A_{OH} = \{(i, j) \in A\colon i \in \mathcal{O}, j \in \mathcal{H}\}$; depot-to-customer arcs, $A_{HD} = \{(i, j) \in A\colon i \in \mathcal{H}, j \in \mathcal{D}\}$; and depot-to-depot arcs, $A_{HH} = \{(i, j) \in A\colon i \in \mathcal{H}, j \in \mathcal{H}\}$. The commodities (types of containers) that move through the network are represented by the set $\mathcal{P}$.

For each supply customer $i \in \mathcal{O}$, the supply of commodity p is noted $o_i^p \geqslant 0$, while for each demand customer $i \in \mathcal{D}$, the demand for commodity p is noted $d_k^p \geqslant 0$. A nonnegative cost c_{ij}^p is incurred for each unit of flow of commodity p moving on arc (i, j). In addition, for each depot $j \in \mathcal{H}$, a nonnegative fixed cost f_j is incurred if the depot is opened.

Let x_{ij}^p represent the flow of commodity p moving on arc (i, j), and y_j the binary location variable that takes value 1 if depot j is opened, and value 0 otherwise. The problem is then formulated as:

$$\text{minimize} \quad \sum_{j\in\mathcal{H}} f_j y_j + \sum_{p\in\mathcal{P}} \Bigg(\sum_{(i,j)\in A_{OH}} c_{ij}^p x_{ij}^p + \sum_{(j,i)\in A_{HD}} c_{ji}^p x_{ji}^p + \sum_{(j,k)\in A_{HH}} c_{jk}^p x_{jk}^p \Bigg) \tag{1}$$

subject to

$$\sum_{j\in\mathcal{H}^+(i)} x_{ij}^p = o_i^p, \quad i \in \mathcal{O}, p \in \mathcal{P}, \tag{2}$$

$$\sum_{j\in\mathcal{H}^-(i)} x_{ji}^p = d_k^p, \quad i \in \mathcal{D}, p \in \mathcal{P}, \tag{3}$$

$$\sum_{i\in\mathcal{D}(j)} x_{ji}^p + \sum_{k\in\mathcal{H}^+(j)} x_{jk}^p - \sum_{i\in\mathcal{O}(j)} x_{ij}^p - \sum_{k\in\mathcal{H}^-(j)} x_{kj}^p = 0, \quad j \in \mathcal{H}, p \in \mathcal{P}, \tag{4}$$

$$x_{ij}^p \leqslant o_i^p y_j, \quad j \in \mathcal{H}, i \in \mathcal{O}(j), p \in \mathcal{P}, \tag{5}$$

$$x_{ji}^{p} \leqslant d_{k}^{p} y_{j}, \quad j \in \mathcal{H}, i \in \mathcal{D}(j), p \in \mathcal{P}, \tag{6}$$

$$x_{ij}^{p} \geqslant 0, \quad (i,j) \in A, p \in \mathcal{P}, \tag{7}$$

$$y_{j} \in \{0,1\}, \quad j \in \mathcal{H}. \tag{8}$$

Constraints (2) and (3) ensure that supply and demand requirements are met, relations (4) correspond to flow conservation constraints at depot sites, while Equations (5) and (6) forbid customer-related movements through closed depots. The analogous constraints for the interdepot flows are redundant if costs satisfy the triangle inequality and are not included in the formulation. The problem has been addressed by dual-ascent heuristics (Crainic and Delorme, 1995), sequential and parallel branch-and-bound (Crainic et al., 1993a; Gendron and Crainic, 1995, 1997; Bourbeau et al., 2000), sequential and parallel tabu search (Crainic et al., 1993c, 1995a, 1995b, 1997; Gendron et al., 1999). Gendron et al. (2003a, 2003b) have studied the capacitated version of the problem and proposed sequential and parallel metaheuristics.

3.1.2 Multicommodity production distribution

The multicommodity production-distribution problem is a simplified version of the previous problem where no interhub movements are considered. In this case, commodities $p \in \mathcal{P}$ may be shipped from their origins $i \in \mathcal{O}$ to their destinations $k \in \mathcal{D}$ either directly or via a consolidation terminal $j \in \mathcal{H}$. (Note that sets $\mathcal{O}$, $\mathcal{D}$, and $\mathcal{H}$ are not necessarily disjoint.) The main decisions addressed by the production-distribution formulations are the number and location of consolidation terminals and the product flow pattern through the system, either directly from the origin to destination or through a consolidation terminal. Particularly relevant for intermodal transportation system design is the representation of the economies of scale in transportation costs associated with concentrated flows (e.g., Croxton et al., 2003, 2006).

The capacity of a consolidation terminal located at site $j \in \mathcal{H}$ is denoted u_j. The transportation cost per unit of flow of commodity p from origin i to destination k transiting through consolidation terminal j is denoted c_{ijk}^{p}, while c_{ik}^{p} stands for the unit transportation cost of direct movements. The corresponding flow routing decision variables are denoted x_{ijk}^{p} and x_{ik}^{p}, respectively. All other notation is the same as in the previous problem. The model is then:

$$\text{minimize} \quad \sum_{j \in \mathcal{H}} f_j y_j + \sum_{p \in \mathcal{P}} \sum_{i \in \mathcal{O}} \sum_{j \in \mathcal{H}} \left(c_{ik}^{p} x_{ik}^{p} + \sum_{k \in \mathcal{D}} c_{ijk}^{p} x_{ijk}^{p} \right) \tag{9}$$

subject to

$$x_{ijk}^{p} \leqslant \min\{o_i^{p}, u_j, d_k^{p}\} y_j, \quad i \in \mathcal{O}, j \in \mathcal{H}, k \in \mathcal{D}, p \in \mathcal{P}, \tag{10}$$

$$\sum_{p \in \mathcal{P}} \sum_{i \in \mathcal{O}} \sum_{k \in \mathcal{D}} x_{ijk}^{p} \leqslant u_j y_j, \quad j \in \mathcal{H}, \tag{11}$$

$$\sum_{k\in\mathcal{D}}\left(x_{ik}^{p}+\sum_{j\in\mathcal{H}}x_{ijk}^{p}\right)\leqslant o_{i}^{p},\quad i\in\mathcal{O},\ p\in\mathcal{P},\tag{12}$$

$$\sum_{i\in\mathcal{O}}\left(x_{ik}^{p}+\sum_{j\in\mathcal{H}}x_{ijk}^{p})=d_{k}^{p},\quad k\in\mathcal{D},\ p\in\mathcal{P},\right.\tag{13}$$

$$y_{j}=0\text{ or }1,\quad j\in\mathcal{H},\tag{14}$$

$$x_{ijk}^{p}\geqslant 0,\ x_{ik}^{p}\geqslant 0,\quad i\in\mathcal{O},\ k\in\mathcal{D},\ j\in\mathcal{H},\ p\in\mathcal{P}.\tag{15}$$

In this formulation, expression (9) minimizes the total cost of opening the consolidation centers and distributing the commodities. Constraints (10) and (11) ensure that only open terminals are used. Relation (11) also enforces the consolidation-terminal capacity constraint. Commodity supplies and demands are enforced by relations (12) and (13), respectively. When demand is specified by origin-to-destination pairs

d_{ik}^{p}: Quantity of commodity p to be transported from origin terminal i to destination terminal k,

the previous formulation is modified by replacing relations (10) with (16) and constraints (12) and (13) with (17):

$$x_{ijk}^{p}\leqslant\min\{u_{j},d_{ik}^{p}\}y_{j},\quad i\in\mathcal{O},\ j\in\mathcal{H},\ k\in\mathcal{D},\ p\in\mathcal{P},\tag{16}$$

$$x_{ik}^{p}+\sum_{j\in\mathcal{H}}x_{ijk}^{p}=d_{ik}^{p},\quad i\in\mathcal{O},\ k\in\mathcal{D},\ p\in\mathcal{P}.\tag{17}$$

Many variants of these formulations may be found in the location literature dealing with practical considerations such as partial, commodity-specific capacities at terminals or handling costs at consolidation terminals (e.g., Aikens, 1985). Piecewise linear functions are often used to represent economies of scale in transportation costs associated with concentrated flows (e.g., Croxton et al., 2003, 2006). Klincewicz (1990) proposed such a formulation for an uncapacitated problem with single-sourcing for each commodity and no direct origin-to-destination movements. When linear transportation costs are assumed either for the origin to consolidation center movements or from the latter to destination terminals, the problem decomposes into concave-cost uncapacitated location problems that can be formulated as uncapacitated plant location problems (with one facility for each piecewise linear segment) for which efficient solution methods exist (e.g., Erlenkotter, 1978). Heuristics that solve a series of these formulations are proposed for the general case. Heuristics combining Lagrangian relaxations and ad-hoc rules based on the logistics characteristics of the system (costs and distances, principally) have also been proposed for the capacitated version of the problem (9)–(15) (Pirkul and Jayaraman, 1996, 1998).

3.1.3 Hub-location models

Consolidation transportation systems are generally structured as *hub-and-spoke* networks (Section 2 and Figure 1). Hub-location models may be used when one must decide simultaneously the location of the hubs (consolidation terminals) and the allocation of regional terminals or customers to the hubs. The location with balancing requirements and the multicommodity production-distribution problems presented in the previous subsections are particular hub-location cases where the allocation of regional terminals or customers to potential hubs does not involve any setup (fixed) cost or has already been decided, respectively. Hub-location formulations appear in many application areas (Campbell, 1994b) and several variants have been proposed and studied (Campbell et al., 2002; Ebery et al., 2000).

The basic model assumes that all traffic passes through two hubs on its route from its origin to its destination, no hub capacities, no direct transport between nonhub terminals, and no fixed costs for establishing a link between a regional and a consolidation terminal. In all hub-based problems, interhub transportation is assumed to be more efficient due to the concentration of flows. Consequently, interhub links are assigned a lower unit cost than links representing the other movements in the system. Most of the notation is similar to that of the previous models. Let us add the following decision-variable definitions:

$y_j =$ 1, if a consolidation terminal is located at site j and 0, otherwise;
$y_{ij} =$ 1, if terminal i is linked to hub j and 0, otherwise;
x^p_{ijlk}: Flow of commodity p with origin i and destination k that passes through hubs j and l, in that order.

The following formulation assumes that exactly P hubs have to be located out of the $|\mathcal{H}|$ potential sites:

$$\begin{aligned}\text{minimize} \quad & \sum_{p\in\mathcal{P}}\Bigg\{\sum_{i\in\mathcal{O}}\sum_{j\in\mathcal{H}} c^p_{ij}y_{ij}\Bigg(\sum_{l\in\mathcal{H}}\sum_{k\in\mathcal{D}} x^p_{ijlk}\Bigg) \\ & \quad + \sum_{l\in\mathcal{H}}\sum_{k\in\mathcal{D}} c^p_{lk}y_{lk}\Bigg(\sum_{i\in\mathcal{O}}\sum_{j\in\mathcal{H}} x^p_{ijlk}\Bigg) \\ & \quad + \sum_{j\in\mathcal{H}}\sum_{l\in\mathcal{H}} c^p_{jl}y_{ij}y_{lk}\Bigg(\sum_{i\in\mathcal{O}}\sum_{k\in\mathcal{D}} x^p_{ijlk}\Bigg)\Bigg\} \qquad (18)\end{aligned}$$

subject to

$$\sum_{j\in\mathcal{H}} y_j = P, \qquad (19)$$

$$\sum_{j\in\mathcal{H}}\sum_{l\in\mathcal{H}} x^p_{ijlk} = d^p_{ik}, \quad i\in\mathcal{O}, k\in\mathcal{D}, p\in\mathcal{P}, \qquad (20)$$

$$y_{ij} \leqslant y_j, \quad i\in\mathcal{O}, j\in\mathcal{H}, \qquad (21)$$

$$y_{lk} \leqslant y_l, \quad k \in \mathcal{D}, l \in \mathcal{H}, \tag{22}$$

$$x^p_{ijlk} \leqslant d^p_{ik} y_j, \quad i \in \mathcal{O}, k \in \mathcal{D}, j, l \in \mathcal{H}, p \in \mathcal{P}, \tag{23}$$

$$x^p_{ijlk} \leqslant d^p_{ik} y_l, \quad i \in \mathcal{O}, k \in \mathcal{D}, j, l \in \mathcal{H}, p \in \mathcal{P}, \tag{24}$$

$$y_j = 0 \text{ or } 1, \quad j \in \mathcal{H}, \tag{25}$$

$$y_{ij} = 0 \text{ or } 1, \quad i \in \mathcal{O}, j \in \mathcal{H}, \tag{26}$$

$$x^p_{ijlk} \geqslant 0, \quad i \in \mathcal{O}, k \in \mathcal{D}, j, l \in \mathcal{H}, p \in \mathcal{P}. \tag{27}$$

The objective function (18) minimizes the total transportation cost of the system. Due to its similarity with the *p-median location* problem, this class of formulations is called the *p-hub median* problem (Campbell, 1994a). Note, however, that despite the name similarity, several properties of the p-median problem do not hold for the p-hub problem (e.g., the assignment of demand nodes to the closest open facility is not optimal for p-hub median problems).

Formulation (18)–(27) was first introduced by O'Kelly (1987). Campbell (1994a) introduced the first linearization mechanism, later refined by Campbell (1996), Skorin-Kapov et al. (1996), and O'Kelly et al. (1996). The mechanism yields a *path*-based, linear mixed-integer formulation:

$$\text{minimize} \quad \sum_{p \in \mathcal{P}} \sum_{i \in \mathcal{O}} \sum_{j \in \mathcal{H}} \sum_{l \in \mathcal{H}} \sum_{k \in \mathcal{D}} c^p_{ijlk} x^p_{ijlk} \tag{28}$$

subject to (19), (20), (23)–(25), and (27), where

c^p_{ijlk}: Unit cost of transportation for commodity p from origin i to destination k with consolidation at hubs j and l, in that order.

The two preceding formulations allow multiple allocations of origin and destination terminals to hubs. To enforce single allocation of terminals to hubs, constraints (23) and (24) are replaced in the first formulation by

$$\sum_{j \in \mathcal{H}} y_{ij} = 1, \quad i \in \mathcal{O}, \tag{29}$$

$$\sum_{l \in \mathcal{H}} y_{lk} = 1, \quad k \in \mathcal{D}, \tag{30}$$

$$\frac{x^p_{ijlk}}{d^p_{ik}} = y_{ij}, \quad i \in \mathcal{O}, k \in \mathcal{D}, j, l \in \mathcal{H}, p \in \mathcal{P}, \tag{31}$$

$$\frac{x^p_{ijlk}}{d^p_{ik}} = y_{lk}, \quad i \in \mathcal{O}, k \in \mathcal{D}, j, l \in \mathcal{H}, p \in \mathcal{P}. \tag{32}$$

Both these constraints and the y_{ij}, y_{lk} variables must be added to the linearized formulation to enforce single terminal allocation.

Both hub-location formulations are difficult to solve. Heuristics have thus been mainly proposed (e.g., O'Kelly, 1987; Aykin, 1990; Klincewicz, 1991;

Campbell, 1996). Metaheuristics have led to improved results: Neural Networks (Smith et al., 1996), Simulated Annealing (Ernst and Krishnamoorthy, 1996), and Tabu Search (Klincewicz, 1992; Skorin-Kapov and Skorin-Kapov, 1994). Lower bounds have also been proposed (e.g., O'Kelly, 1992a; O'Kelly et al., 1995; Skorin-Kapov et al., 1996), as well as a number of methods that find the optimal solution to problem instances of limited size (Skorin-Kapov et al., 1996; O'Kelly et al., 1996; Ernst and Krishnamoorthy, 1996).

A number of important extensions to these models may be formulated. Fixed cost terms $\sum_{j\in\mathcal{H}} f_j y_j$ added to the objective function (and dropping constraint (19) on the number of hubs to open) may be used to represent the costs of opening consolidation centers (O'Kelly, 1992b). Klincewicz (1996) proposed dual-based heuristics for this formulation, while Addinnour-Helm and Venkataramanan (1998) developed a genetic search metaheuristic and a branch-and-bound algorithm. Limited-sized problem instances have been addressed.

Additional extensions consider fixed costs for allocating origin and destination terminals to hubs, hub capacities, more complex routing patterns (e.g., more than two hubs), minimum levels of traffic in order to allow a terminal-to-hub connection, etc., and are better able to capture the complexity of transportation systems. Such hub network design formulations are very difficult to solve, however, and relatively little research had been conducted (e.g., Aykin, 1994, 1995a, 1995b; Kuby and Gray, 1993; Jaillet et al., 1996). Significant research is required in this domain.

3.2 Service network design

Service network design formulations are used to build a transportation (or load) plan to ensure that the system operates efficiently, serves demand, and ensures the profitability of the firm. The physical infrastructure (e.g., the terminal locations) and the available resources are fixed for these problems. Service network design problems address the system-wide planning of operations to decide the selection, routing, and scheduling of services, the consolidation activities in terminals, and the routing of freight of each particular demand through the physical and service network of the company. The goal is cost-efficient operation together with timely and reliable delivery of demand according to customer specifications and the targets of the carrier.

Service network design problems are difficult due to the strong interactions among system components and decisions and the corresponding tradeoffs between operating costs and service levels. To illustrate, consider the routing of a shipment between two terminals of the intermodal transportation system illustrated in Figure 1. (Assume, for simplicity, that each link corresponds to a service.) A shipment that originates at terminal 4 with destination terminal 8 may be routed according to a number of strategies, including (1) direct through its designated hub B; (2) indirect through hub A or terminal 5 and

then through B; (3) indirect through hubs A and C. The number of strategies increases rapidly when one considers additional factors such as the type of operation (e.g., consolidation or simple cargo transfer from one service to another) or consolidation policy.

Which alternative is the "best"? Each has its own cost and time measures that result from the characteristics of each terminal and service, as well as from the routing of all other shipments. Thus, for example, strategies based on re-consolidation and routing through intermediate terminals may be more efficient when direct services are offered rarely due to low levels of traffic demand. Such strategies would probably result in higher equipment utilization and lower waiting times at the original terminal; hence, in a more rapid service for the customer. The same decision would also result, however, in additional unloading, consolidation, and loading operations, creating larger delays and higher congestion levels at terminals, as well as a decrease in the total reliability of the shipment. Alternatively, offering direct service (introducing the service or increasing its frequency) would imply faster and more reliable service for the corresponding traffic and a decrease in the level of congestion at some terminals, but at the expense of additional resources, thus increasing the direct costs of the system.

Service network design thus integrates two types of major decisions. The first is to determine the service network, that is, to *select* the routes – origin and destination terminals, physical route and intermediate stops – on which services will be offered and the characteristics of each service: mode, *frequency* or *schedule*, etc. The second major type of decision is to determine the routing of demand, that is, the *itineraries* used to move the flow of each demand: services and terminals used, operations performed in these terminals, etc. The service network specifies the movements through space and time of the vehicles and convoys of the various modes considered. Operating rules indicating, for example, how cargo and vehicles may be sorted and consolidated are sometimes specified at particular terminals and become part of the service network (this is the case, in particular, for rail carriers). The itineraries used to move freight from origins to destinations determine the flows on the services and through the terminals of the service network.

Minimization of the total operating costs is the primary optimization criterion, reflecting the traditional objectives of freight carriers and expectation of customers to "get there fast at lowest possible cost". Increasingly, however, customers not only expect low rates, but also require a high-quality service, measured by speed, flexibility, and reliability. Service performance measures modeled, in most cases, by delays incurred by freight and vehicles or by the respect of predefined performance targets are then added to the objective function of the network optimization formulation. The resulting generalized cost function thus captures the tradeoffs between operating costs and service quality.

Several efforts have been directed toward the formulation of service network design models. Reviews are presented by Assad (1980), Crainic (1988,

2000, 2003), Delorme et al. (1988), and Cordeau et al. (1998). Most of these contributions aim specific carrier types and modes of transportation (rail, less-than-truckload trucking, navigation, etc.) and are thus covered in more depth in other chapters of this book. In the following, we briefly present basic modeling approaches and examine a number of contributions aimed at intermodal transportation.

One may distinguish between static and time-dependent service network design formulations. The former assume that demand does not vary during the planning period that is considered. The time dimension of the service network is then implicitly considered through the definition of services and interservice operations at terminals. Time-dependent formulations include an explicit representation of movements in time and usually target the planning of *schedules* to support decisions related to *when* services depart. Most service network design models proposed in the literature take the form of *deterministic, fixed cost, capacitated, multicommodity network design* formulations.

3.2.1 Static service network design

Service network design assumes a physical network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ on which the carrier may move its vehicles. On this network, the carrier operates (or makes use of) a number of terminals, with various characteristics and functionalities (loading/unloading stations, regional terminals with limited sorting capabilities, consolidation hubs, etc.). Let nodes in $\mathcal{H} \subseteq \mathcal{N}$ correspond to locations where the terminals are situated and assume, for simplicity, that all terminals can perform all operations.

The *service network* specifies the transportation services that could be offered. Each *service* $s \in \mathcal{S}$ is characterized by its (1) mode, which may represent either a specific transportation mode (e.g., rail and truck services may belong to the same service network), or a particular combination of traction and service type (e.g., fast navigation lines by large container ships providing shuttle service between South-East Asia and the west coast of North America); (2) route, defined as a path in $\mathcal{A}$, from its origin terminal to its destination terminal, with intermediary terminals where the service stops and work may be performed (e.g., a navigation line that stops at pre-defined ports for unloading and loading containers); (3) capacity, which may be measured in load weight or volume, number of containers, number of vehicles (when convoys are used to move several vehicles simultaneously), or a combination thereof; (4) service class that indicates characteristics such as preferred traffic or restrictions, speed and priority, etc. To design the service network thus means to decide what service to include in the transportation plan such that the demands and the objectives of the carrier are satisfied. When a service may be operated repeatedly during the planning period (e.g., several similar ships departing during the month and visiting the same ports in the same order), the design must also determine the frequency of each service.

Crainic and Rousseau (1986) proposed a multimodal multicommodity path-flow service network design formulation. In their model, a commodity $p \in \mathcal{P}$ is

defined as a triplet (origin, destination, type of product (or vehicle)) and traffic moves according to *itineraries*. An itinerary $l \in \mathcal{L}^p$ for commodity p specifies the service path used to move (part of) the corresponding demand: the origin and destination terminals, the intermediary terminals where operations (e.g., consolidation and transfer) are to be performed, and the sequence of services between each pair of consecutive terminals where work is performed. The demand of product p is denoted d_p. Flow routing decisions are then represented by decision variables h_l^p indicating the volume of product p moved by using its itinerary $l \in \mathcal{L}^p$. *Service frequency* decision variables y_s, $s \in \mathcal{S}$, define the level of service offered, i.e., how often each service is run during the planning period.

Let $y = \{y_s\}$ and $h = \{h_l^p\}$ be the decision-variable vectors. The model minimizes the total generalized system cost, while satisfying the demand for transportation and the service standards:

$$\text{minimize} \quad \sum_{s\in\mathcal{S}} \mathcal{F}_s(y) + \sum_{p\in\mathcal{P}} \sum_{l\in\mathcal{L}^p} \mathcal{C}_l^p(y,h) + \Theta(y,h) \tag{33}$$

$$\text{subject to} \quad \sum_{l\in\mathcal{L}^p} h_l^p = d_p, \quad p \in \mathcal{P}, \tag{34}$$

$$y_s \geqslant 0 \text{ and integer}, \quad s \in \mathcal{S}, \tag{35}$$

$$h_l^p \geqslant 0, \quad l \in \mathcal{L}, p \in \mathcal{P}, \tag{36}$$

where

$\mathcal{F}_s(y)$: Total cost of operating service s;
$\mathcal{C}_l^p(y,h)$: Total cost of moving (part of) product p demand by using its itinerary l;
$\Theta(y,h)$: Penalty terms capturing various relations and restrictions, such as limited service capacity.

The objective function defines the total system cost and includes the total cost of operating a service network at given frequencies, the total cost of moving freight by using the selected itineraries for each commodity, as well as a number of terms translating operational and service restrictions, such as facility and service capacities and on-time delivery targets, into monetary values. This third term models, for example, service capacity restrictions as utilization targets, over-assignment of traffic being permitted at the expense of additional costs and delays. The introduction of the third term aims to enhance the capability of the model to identify tradeoffs between the cost of increasing the level of service (service frequency) and the cost of routing freight on less interesting itineraries.

The objective function computes a generalized cost, in the sense that it may include various productivity measures related to terminal and transportation operations. In addition to the actual costs of performing the operations, one may thus explicitly consider the costs, delays, and other performance measures

related to the quality and reliability of the service offered. Nonlinear congestion functions were thus used to reflect the increasingly larger delays that result when facilities of limited capacity (e.g., consolidation terminals) must serve a growing volume of traffic brought by vehicles of different services carrying freight for different products on various itineraries. The notation aims to convey this utilization of more general functional forms.

The model was adapted and applied to rail, long-haul LTL trucking (railway transportation of trailers was included as a possible mode), and express letter services (e.g., Crainic et al., 1984; Roy and Delorme, 1989; Roy and Crainic, 1992). The original solution method described by Crainic and Rousseau (1986) combines a heuristic that iteratively decreases frequencies from initial high values, and a convex network optimization procedure to distribute the freight. The latter makes use of column generation to create itineraries and descent procedures to optimize the flow distribution.

Most service network design models proposed in the literature use arc-flow-based formulations and focus on the selection decisions. This results into {0, 1} network design formulations (e.g., the liner service design models presented by Christiansen et al., 2007). Minimum thresholds are imposed in some cases on the volume of traffic required to allow the selection of a transportation service (e.g., the load planning model for LTL motor carriers introduced by Powell (1986), Powell and Sheffi (1983, 1986, 1989), Braklow et al. (1992)).

3.2.2 Design of postal services

The design of postal services, regular and those dedicated to express letter and package delivery, yields particularly large and complex service network design formulations. These applications involve air and land long-haul transportation, as well as local pick up and delivery operations. Land long-haul transportation is often performed by using trucks, although rail is also used in some countries. It should be noted that the expansion of high-speed railway networks is making rail increasingly attractive even for express services. The air mode is both expensive and efficient in carrying relatively large loads over long distances. Due to the high costs, efforts are dedicated to the optimization of available capacity and the reduction of the number of aircrafts used. This translates into a relatively large number of alternatives in terms of aircraft types, capacities, and costs which adds to the complexity of the problem.

Kuby and Gray (1993) developed an early model for the design of the network of an express package delivery firm. It is a path-based {0, 1} network design model, where multistop aircraft routes must be selected in and out of a given hub. Paths were generated a priori and the model was solved with a standard mixed-integer package. Analyses illustrated the cost effectiveness of a design with multiple stops over a pure hub-and-spoke network.

Kim et al. (1999) and Armacost et al. (2002) propose a more comprehensive model for the design of the multimodal version of the problem where both air and ground vehicles are considered (Barnhart and Schneur, 1996, address a simplified, single-hub version of the problem). In this problem, several hubs

and aircraft types are considered, while trucks may perform pickup and delivery activities, as well as transportation over limited distances. The focus is on the design of the overnight air services for next-day delivery.

Two types of terminals are considered in the model: *gateways* where packages enter and exit the air network and hubs where packages are unloaded from in-coming airplanes, sorted, and loaded into an out-going aircraft for delivery to their destination gateways. Aircrafts may fly nonstop between gateways and hubs or stop at one gateway en-route. *Time window* restrictions on pickup and delivery times at gateways as well as on the sorting periods at hubs limit the number of stops and constrain the design. Aircrafts of different types are available in limited numbers. Each aircraft type has an operating cost and a capacity, as well as operating characteristics (type of engine, flying range, speed, etc.) that determine the routes it can fly. The objective is to design the minimum cost set of routes, aircraft assignments to these routes, and package flows to satisfy demand and level-of-service objectives, while complying with the operating parameters of the terminals (e.g., capacity), aircraft type (e.g., capacity, range, speed), and aircraft fleet (number of planes available).

A commodity $p \in \mathcal{P}$ represents packages to be moved from an origin gateway $o(p)$ to a destination gateway $d(p)$. Let d^p indicate the corresponding demand, $\mathcal{H}$ the set of hubs, and $\{o(p), d(p), p \in \mathcal{P}\}$ the set of origin and destination gateways. Let $\mathcal{F}$ be the set of aircraft types available and $\mathcal{R}^f$ the set of routes (sequence of gateways starting or ending at a hub) that may be flown by aircrafts of type $f \in \mathcal{F}$. The routes become the flight arcs $\mathcal{A}$ linking the hub and gateway nodes $\mathcal{N}$ of the system. The cost of flying route $r \in \mathcal{R}^f$ with aircrafts of type $f \in \mathcal{F}$ is denoted c_r^f, while u_r^f gives its capacity. Costs associated to package flows are not significant compared to those of operating the air fleets and are thus not included in the formulation. The number of available aircrafts of type f is n_f and the landing capacity of hub $h \in \mathcal{H}$ is a_h. Three indicators are also defined: (1) $\delta_{ij}^{rf} = 1$ when flight arc (i, j) is included in route r flown by aircraft type f and 0, otherwise; (2) $\beta_i^r = 1$ (-1) if node i is the origin (destination) gateway of route r and 0, otherwise; and (3) $\delta_r^h = 1$ if hub h is included in route r and 0, otherwise.

Two types of decision variables are defined: y_r^f indicates the number of times route $r \in \mathcal{R}^f$ is flown by aircrafts of type $f \in \mathcal{F}$; continuous variable x_{ij}^p stands for the amount of product p on the air link $(i, j) \in \mathcal{A}$. The basic frequency service network design formulation is then:

$$\text{minimize} \quad \sum_{f \in \mathcal{F}} \sum_{r \in \mathcal{R}^f} c_r^f y_r^f \tag{37}$$

subject to

$$\sum_{j \in \mathcal{N}} x_{ij}^p - \sum_{j \in \mathcal{N}} x_{ji}^p = \begin{cases} d^p & \text{if } i = o(p), \\ -d^p & \text{if } i = d(p), \\ 0 & \text{otherwise,} \end{cases} \quad i \in \mathcal{N},\ p \in \mathcal{P}, \tag{38}$$

$$\sum_{p\in\mathcal{P}} x_{ij}^{p} \leqslant \sum_{f\in\mathcal{F}} \sum_{r\in\mathcal{R}^f} \delta_{ij}^{rf} u_r^f y_r^f, \quad (i,j)\in\mathcal{A}, \tag{39}$$

$$\sum_{r\in\mathcal{R}^f} y_r^f \leqslant n_f, \quad f\in\mathcal{F}, \tag{40}$$

$$\sum_{r\in\mathcal{R}^f} \beta_i^r y_r^f = 0, \quad i\in\mathcal{N}, f\in\mathcal{F}, \tag{41}$$

$$\sum_{r\in\mathcal{R}^f} \delta_r^h y_r^f \leqslant a_h, \quad h\in\mathcal{H}, \tag{42}$$

$$y_r^f \geqslant 0 \text{ and integer}, \quad r\in\mathcal{R}^f, f\in\mathcal{F}, \tag{43}$$

$$x_{ij}^{p} \geqslant 0, \quad (i,j)\in\mathcal{A}, p\in\mathcal{P}. \tag{44}$$

Constraints (38) enforce flow conservation for each product, while the forcing constraints (39) restrict the flow on each fly arc to the capacity of the aircrafts flown on that route. Constraints (40) and (42) enforce the number of available aircrafts of each type and the landing capacities at hubs, respectively. Constraints (41) are the aircraft balance restrictions: the number of aircrafts of each given type landing at a location (delivery routes) must equal the number taking off from that same location (pick up routes).

In Kim et al. (1999), the authors examine arc, path, and tree-based formulations of this model, and select the latter since it yields a problem of significantly reduced size compared to the two others. The authors solve the linear relaxation of the resulting formulation by combining heuristics, which further reduce the size of the problem, cut-set inequalities to strengthen the relaxation, and column generation to gradually generate a good set of route variables. Branch-and-bound is then used to obtain an integer solution. A different methodological approach is presented by Armacost et al. (2002). The authors transform the problem formulation by defining new composite variables that combine the original air service frequency and the package flow variables to represent possible air routes from gateway to hub with minimal but sufficient capacity to transport the required volume. The composite-variable formulation includes constraints that force the selection of at least one composite variable for each gateway–hub connection. The authors take advantage of the fact that air routes pass at most through two gateways and thus the number of composite variables, which allow to capture multiple aircraft routes with a single variable, is relatively limited. Moreover, the composite route variables implicitly account for the flow distribution and thus yield a pure design formulation for which stronger bounds and thus more efficient solution methods may be derived. The methodology has been implemented at a major US express package delivery company with significant success (Armacost et al., 2004). The model is used on a continuous basis for the planning of next-day operations as well as for what-if scenario analyses. Such results emphasize the need to continue to explore the network design formulations for new insights and more efficient solution methods.

The reorganization of the German postal services belongs to the same problem class, albeit on a more comprehensive scale. Grünert and Sebastian (2000) (see also Grünert et al., 1999; Buedenbender et al., 2000) decompose the problem into several subproblems of manageable proportions: the optimization of the overnight airmail network, the design of the ground-feeding and delivery transportation system, and the scheduling of operations. Vehicle routing models and techniques, as well as a discrete, time-dependent network design formulation (see Section 3.2.3), are proposed for the routing and scheduling tasks. The air network design formulation is further decomposed into a direct flight problem and a hub system problem; both subproblems are fixed cost, multicommodity, capacitated network design formulations with side constraints. To optimize these formulations, the authors propose combinations of classical heuristics, tabu search and evolutionary metaheuristics, and exact mathematical programming methods (branch-and-bound). A decision support system integrates the models and associated solution methods, as well as the tools required to handle the data, models, and methods, and to assist in the decision process.

3.2.3 Time-dependent service network design

When schedules are contemplated, a *time* dimension must be explicitly introduced into the formulation. This is usually achieved by representing the operations of the system over a certain number of *time periods* by using a *space–time* network. The representation of the physical network is replicated in each period. Starting from its origin in a given period, a service arrives (and leaves, in the case of intermediary stops) at a later period at other terminals. Services thus generate temporal service links between different terminals at different time periods. Temporal links that connect two representations of the same terminal at two different time periods may represent the time required by terminal activities or the freight waiting for the next departure. The costs associated with the arcs of these networks are similar to those used in the static formulations of the previous subsections. Additional arcs may be used to capture penalties for arriving too early or too late.

There are again two types of decision variables. Integer design variables are associated with each service. Restricted to $\{0, 1\}$ values, these variables indicate whether or not the service leaves at the specified time. When several departures may take place in the same time period, general (nonnegative) integer variables must be used. (Note that one can always use $\{0, 1\}$ variables only by making the time periods appropriately small.) Continuous variables are used to represent the distribution of the freight flows through this service network.

The resulting formulations are network design models similar to those presented in the previous subsections but on significantly larger networks due to the time dimension. The sheer size of the space–time network, as well as the additional constraints usually required by the time dimension, makes this class of problems more difficult to solve than static versions. A limited number of contributions have thus been reported to date. Most address

the problem in the context of a single carrier and propose heuristic solution methods (e.g., Farvolden and Powell, 1991, 1994; Farvolden et al., 1992; Equi et al., 1997).

Different formulations are being proposed to address the recent trends in the development and operation policy of rail intermodal networks. One such trend is the introduction of booking systems that force customers to book in advance a precise number of container space slots on a given day and train service. To operate such systems, intermodal rail carriers enforce regular and cyclically scheduled services. Moreover, in order to decrease operating costs (by simplifying operations and reducing equipment handling) and to increase the utilization of their assets, rail cars and engines, mainly, the characteristics of services in terms of composition, capacity, and so on, are supposed to stay the same for all the time the schedule is valid. This starts to be known as a "full-asset-utilization" operating policy.

Not much work has yet been done with respect to these new operation characteristics, which may be observed both in North America and Europe. Pedersen et al. (2006) present one of the first contributions. The authors notice that a "full-asset-utilization" operating policy requires that the asset circulation issue be integrated into the service network design model. They represent this requirement by enforcing the condition that at each node of the representation, the (integer) design flow be balanced. The authors propose two formulations (a link and a cycle-based one) and a tabu search-based metaheuristic that appears to efficiently yield good quality solutions. Much more work is required in this area, however.

3.3 Port dimensioning

Important strategic planning issues for the design of container port terminals are related to the number of berths, the size of storage space, and the number of pieces of various equipment to install. Port dimensioning issues require a trade-off between the amount of investment and the level of customer service. For example, as the number of berths increases, the turnaround (waiting) time of vessels decreases. Because of the high investment and operating cost of container ships, delays experienced at a port generate high costs to the ship's operators and cause delayed arrivals at successive ports creating serious downstream operational problems. Thus, availability of empty berths at a port is a key issue when operators determine container terminals for their ships. Studies aimed at determining the optimal number of berths usually consider the costs related to the turnaround time of ships, the berth construction cost, and the berth operation cost.

Quay cranes are the most expensive (5–10 million dollars per quay crane) handling equipment in port container terminals. Determining the number of quay cranes to be installed is thus another major container–port design issue. In addressing this issue, one must consider that as the number of quay cranes per berth increases, the throughput rate per berth increases as well, but at a

declining rate due to interference between adjacent quay cranes. Determining the number of pieces of other types of equipment, such as transporters, yard cranes, and so on, are also important port dimensioning issues.

Storage space is another critical resource in port container terminals, especially in major Asian hub ports such as Singapore, Hong Kong, Busan, Kobe, and Kaohsiung. As the storage space becomes smaller, the storage stacks become higher, which results in lower productivity of the transfer operations in the yard. Thus, there is a trade-off relationship between the investment in storage space and the productivity of the transfer operations in the yard (see Section 5).

Very few optimization-based models have been proposed for these issues, however. Descriptive models, mainly based on queuing theory and probabilistic approaches, have been used in most cases. The goal was to evaluate the performance of particular components of container terminals under various hypotheses regarding the number, type and, eventually, combination of resources, e.g., Fratar et al. (1960), Miller (1971), Wanhill (1974), and Daganzo (1990) for the number of berths; Griffiths (1976) and Daganzo (1989a) for the number of quay cranes; Schonfeld and Sharafeldien (1985) for the number of quay cranes and berths; van Hee and Wijbrands (1988), Kozan (1997), and Kim and Kim (2002) for combinations of different types of resources including yard cranes, storage slots, and yard trailers. Two exceptions to this trend are nonlinear programming model proposed by Noritake and Kimura (1990) for berth dimensioning and the maximum flow network model introduced by Vis et al. (2001) to determine the minimum number of automated guided vehicles for completing a given set of delivery tasks.

Simulation models have also been proposed to evaluate seaport container terminal design (e.g., number and layout of various resources), operation policies (e.g., working hours), operation characteristics (e.g., travel and handling times of various equipments) and general performance: Ramani (1996), Lai and Leung (2000), Nam et al. (2002), etc. A different simulation approach was proposed by Alessandri et al. (2006), which represent containers and their movements in a terminal as a network of queues. Discrete-time equations are then used to describe the dynamic evolution of the system, where control variables represent the utilization of the terminal resources. This model yields an optimization problem that aims to minimize the transfer delays in the terminal. The problem is stated as an optimal control problem and is addressed by a receding-horizon solution strategy.

4 Container fleet management

The need for freight carriers to move empty vehicles follows from the differences in demand and supply for each commodity observed at most locations, resulting in an accumulation of empty vehicles in regions where they are not

needed and in deficits of vehicles in other regions that require them. Vehicles must then be moved empty, or additional loads must be found, in order to bring them where they are needed to satisfy known and forecast demand in the following planning periods. This operation is known as *repositioning* or *empty balancing* and is a major component of what is known as *fleet management*. In its most general form, fleet management covers the whole range of planning and management issues from procurement of power units and vehicles to vehicle dispatch and scheduling of crews and maintenance operations. Often, however, the term designates a somewhat restricted set of activities: allocation of vehicles to customer requests and repositioning of empty vehicles. We follow this definition is this section.

Fleet management belongs to what is usually called the operational level of planning. Most system elements, demand, travel and handling times, and so on, vary with time. Most strategic and tactical planning models do not explicitly account for these variations. At the operational planning level, however, the time-dependency of data, decisions, and operations must be explicitly addressed, including the representation of the outcome of current decisions on the future state of the system. The *time-dependent* (*dynamic*) aspect of operations is further compounded by the *stochasticity* inherent to most systems, that is, by the set of uncertainties that are characteristic of real-life operations and management activities. If traffic is slower than predicted, vehicles may arrive late at customers' locations or at the terminal. Forecast customer requests for empty containers may not materialize while unexpected demands may have to be satisfied. The planned supplies of empty vehicles at depots may thus be unsettled and additional empty movements may have to be performed. Increasingly, these characteristics are reflected in the models and methods aimed at operational planning and management issues, as illustrated by the fleet management models of this section.

Moving vehicles empty does not directly contribute to the profit of the firm but it is essential to its continuing operations. Consequently, one attempts to minimize empty movements within the limits imposed by the demand and service requirements. Early models were based on transportation formulations or, most often, deterministic time-dependent transshipment network models (e.g., White, 1972, and Ermol'ev et al., 1976, for container fleet management). Details on the early approaches may be found in Dejax and Crainic (1987). Models that explicitly consider uncertainties in empty vehicle distribution started to be proposed in the mid-1980s (e.g., Jordan and Turnquist, 1983). There is now a significant body of work in this area, mostly dedicated to rail and motor carrier issues (see the reviews by Crainic (2003), Cordeau et al. (1998), Powell (2003), Powell and Topaloglu (2003, 2005), Powell et al. (2007)).

Few efforts were dedicated to container fleet management issues. Crainic et al. (1993b) proposed a series of models for the allocation and management of a heterogeneous fleet of containers where loaded movements are exogenously accepted. Cheung and Chen (1998) focused on the single-commodity container allocation problem for liner regular ocean navigation lines operators.

Powell and Carvalho (1998) addressed the problem of the combined optimization of containers and flatcars for rail intermodal operations.

The container fleet management problem addressed by Crainic et al. (1993b) was cast in the context of the land transport part, as described in Section 3.1.1. The land container distribution system is composed of a number of port terminals, in-land depots, and customer locations. Ships arrive in ports carrying loaded and empty containers of various types and dimensions. Loaded containers are delivered to their destinations, using rail and truck under various types of contracts, while empty containers are available for delivery to customers in the vicinity of the port or for repositioning. Customers receiving loaded containers unload them and signal that they may be picked up and transported to a designated terminal (port or in-land). Similarly, customers that require empty containers of specific types for future shipments receive them from an in-land or a port terminal. If the number of containers of a given type requested by the customer is not available for delivery within the time window specified by the customer, the company can lease (or borrow from partner companies) containers or substitute certain other types of containers. The empty containers available for distribution to customers thus come from four sources: customers that unloaded their goods, world-wide repositioning, leasing, and substitution. Containers are repositioned through two mechanisms: (1) the balancing movements determined by the strategic/tactical planning of the operations (Section 3.1.1), and (2) the allocation decisions concerning the depots to which the containers that become empty at customer locations are to be sent. The problem is dynamic, in the sense that demands vary in time and transport and customer operations may both require more than one period (day). It is stochastic due to variations in customer demands, including the possibility of demands from "unknown" customers, variations in the time required by customers to unload and return containers, and damage to containers (travel time variations were not considered in the chapter).

The formulation is defined for a planning horizon discretized in T periods: $t = 1, 2, \ldots, T$. Let $\mathcal{H}$ be the set of ports and $\mathcal{D}$ the set of inland depots. The sources of randomness considered in the formulation are:

d_i^p: Demand of containers of type p for demand customer $i \in \mathcal{I}^p$; The containers must be delivered within a time window Δ_i;

d_h^{pt}: Demand of empty containers of type p at port $h \in \mathcal{H}$ in period t for export;

o_s^{pt}: Supply of empty containers of type p released by supply customer s at time t; The set of these customers is denoted $\mathcal{S}^{pt}$;

o_h^{pt}: Supply of empty containers of type p arriving at port h in period t.

The parameters of the problem are:

$\mathcal{J}_i^{pt}$: Set of depots j that may send a shipment of containers of type p that would arrive in period t at customer i;

$\mathcal{I}_j^{pt}$: Set of customer requests that may be served (i.e., he containers of type p will arrive within Δ_i) by a shipment starting from depot $j \in \mathcal{D} \cup \mathcal{H}$ in period t;

c_{ji}^{pt}: Unit transportation cost for a container of type p from a depot j to demand customer $i \in \mathcal{I}_j^{pt}$;

$\mathcal{S}_j^{pt}$: Set of supply customers from where containers of type p may reach terminal j in period t;

c_{sj}^{pt}: Unit transportation cost for a container of type p from supply customer $s \in \mathcal{S}_j^{pt}$ to a depot j;

c_{jk}^{pt}: Unit transportation cost for a container of type p between depots $j, k \in \mathcal{D} \cup \mathcal{H}$;

c_{pr}: 1/(Number of containers of type p needed to replace one container of type r);

c_j^{prt}: Unit cost at depot j in period t to substitute a container of type p for a container of type r;

c_j^{pt}: Unit holding cost for a container of type p at depot j in period t;

$\bar{c}_j^{pt}$: Cost of leasing (or borrowing) a container of type p at depot j in period t; It also represents the penalty cost at port h when d_h^{pt} cannot be satisfied.

With decision variables:

x_{ji}^{pt}: Number of equivalent containers of type p allocated in period t from depot j to customer i;

x_j^{pt}: Number of containers of type p allocated as type p in period t from depot j;

x_j^{prt}: Number of containers of type p substituted for containers of type r in period t at depot j;

x_{sj}^{pt}: Number of containers of type p picked up at customer s in period t and delivered to depot j in period $\bar{t} \geqslant t$;

w_j^{pt}: Number of containers of type p in stock at depot j at the end of period t;

w_{jk}^{pt}: Number of containers of type p repositioned from depot j to depot k in period t;

b_j^{pt}: Number of containers of type p rented at depot j in period t;

the model minimizes the expected total operating cost, including substitutions and stockouts, over a multiperiod planning horizon

$$E\Bigg[\sum_{t}\sum_{p\in\mathcal{P}}\sum_{j\in\mathcal{H}\cup\mathcal{D}}\Bigg(\sum_{s\in\mathcal{S}_j^{pt}} c_{sj}^{pt}x_{sj}^{pt}+\sum_{i\in\mathcal{I}_j^{pt}} c_{ji}^{pt}x_{ji}^{pt}+\sum_{k\in\mathcal{H}\cup\mathcal{D}} c_{jk}^{pt}w_{jk}^{pt} + \sum_{r\in\mathcal{P}} c_j^{prt}x_j^{prt}+c_j^{pt}w_j^{pt}+\bar{c}_j^{pt}b_j^{pt}\Bigg)\Bigg] \tag{45}$$

subject to nonnegativity constraints on all variables, bounds on balancing flows, and constraints representing the dynamics of the system:

$$\sum_{t\in\Delta_i}\sum_{j\in\mathcal{J}_i^{pt}} x_{ji}^{pt}-d_i^p=0,\quad i\in\mathcal{I}^p, \tag{46}$$

$$\sum_{\bar{t}\geqslant t}\sum_{j\in\mathcal{S}_{jp\bar{t}}} x_{sj}^{p\bar{t}}-o_s^{pt}=0,\quad s\in\mathcal{S}^{pt},\ p\in\mathcal{P},\ t=1,2,\ldots,T, \tag{47}$$

$$x_j^{pt}+\sum_{r\in\mathcal{P}} c_{pr}x_j^{prt}-\sum_{i\in\mathcal{I}_j^{pt}} x_{ji}^{pt}=0,\quad j\in\mathcal{D}\cup\mathcal{H},\ p\in\mathcal{P},\ t=1,2,\ldots,T, \tag{48}$$

$$w_j^{pt}+x_j^{pt}+x_j^{prt}+\sum_{k\in\mathcal{H}\cup\mathcal{D}} w_{jk}^{pt}-w_j^{pt-1}-\sum_{s\in\mathcal{S}_j^{pt}} x_{sj}^{pt}-\sum_{\bar{t}\leqslant t}\sum_{k\in\mathcal{H}\cup\mathcal{D}} w_{kj}^{p\bar{t}}-b_j^{pt}=0,\quad j\in\mathcal{D},\ p\in\mathcal{P},\ t=1,2,\ldots,T, \tag{49}$$

$$w_h^{pt}+x_h^{pt}+x_h^{prt}+\sum_{k\in\mathcal{H}\cup\mathcal{D}} w_{hk}^{pt}+d_h^{pt}-w_{hk}^{pt}-\sum_{s\in\mathcal{S}_h^{pt}} x_{sh}^{pt}-\sum_{\bar{t}\leqslant t}\sum_{k\in\mathcal{H}\cup\mathcal{D}} w_{hk}^{p\bar{t}}-o_h^{pt}-b_h^{pt}=0,\quad h\in\mathcal{H},\ p\in\mathcal{P},\ t=1,2,\ldots,T, \tag{50}$$

where constraints (46) and (47) enforce demand and supply conditions at customers, Equations (49) and (50) are the flow conservation conditions at in-land depots and ports, respectively, while Equations (48) represent the allocation accounting between true and substitution containers at depots. The authors presented single and multicommodity deterministic formulations and a two-stage, restricted-recourse single commodity, stochastic model.

The stochastic and dynamic empty container allocation model of Cheung and Chen (1998) addresses the problem from the point of view of a container liner company that offers regular service lines between a given number of ports. Ships transport loaded containers that bring a profit and, space permitting, empty containers to reposition them in order to be able to satisfy future forecast demand. It is assumed that only one ship travels between two ports at each time period and that demand for loaded containers between ports does not exceed the capacity of the ship. Only one product – empty containers – is considered. The schedule of ships linking the ports in $\mathcal{H}$ is assumed fixed for the planning horizon $t = 1, 2, \ldots, T$. Similarly to the formulations of Crainic et al. (1993b), one supposes that the model is to be run into a rolling horizon mode: the model gives a solution (that accounts for the impact of today's decisions on the future state of the system), the suggestion solution for the first period is implemented and then, in the next period, the model is run again using the up-dated information.

Three sources of randomness are considered in the model:

κ_{ij}^t: residual capacity for empty containers on the ship traveling from port $i \in \mathcal{H}$ to port $j \in \mathcal{H}$, leaving at time t;
δ_i^t: Demand for containers at port i at period t;
σ_i^t: Supply of containers at port i at period t (before unloading from ships).

The parameters of the problem are:

c_{ij}^t: Unit transportation cost from port $i \in \mathcal{H}$ to port $j \in \mathcal{H}$, leaving at time t;
l_i^t: Unit cost for loading a container on a ship at port i at period t;
u_i^t: Unit cost for unloading a container from on a ship at port i at period t;
c_i^t: Unit holding cost for a container at port i in period t;
$\bar{c}_i^t$: Unit cost of leasing (or borrowing) at port i in period t;
R_i^t: Unit revenue from satisfying demand at port i at period t;
τ_{ij}: Transportation time from port i to port j;

and the decision variables:

$x_i^{l,t}$: Number of containers loaded at port i in period t for any destination;
$x_i^{u,t}$: Number of containers unloaded at port i in period t from any other port;
$x_{ij}^{c,t}$: Number of containers to be repositioned from port i to port j departing in period t;
$x_{ij}^{p,t}$: Number of containers currently on the ship, in port i at period t, with destination port j;
$x_i^{h,t}$: Number of containers stored at port i at period t;

$x_i^{r,t}$: Number of leased containers at port i and period t used to meet demand;

$x_i^{d,t}$: Number of owned containers at port i and period t used to meet demand.

All decision variables are nonnegative. The objective is to minimize the expected total cost while maximizing the revenue from satisfying demand:

$$\text{minimize} \quad E\Bigg[\sum_t \sum_{i\in\mathcal{H}}\Bigg(l_i^t x_i^{l,t} + u_i^t x_i^{u,t} + c_i^t x_i^{h,t} + \bar{c}_i^t x_i^{r,t} + \sum_{j\in\mathcal{H}} c_{ij}^t x_{ij}^{c,t} - R_i^t \delta_i^t\Bigg)\Bigg] \tag{51}$$

while the following relations define the system dynamics and constraints of the model:

$$\sum_{k\in\mathcal{H}} x_{ki}^{c,t-\tau_{ij}} - x_i^{u,t} - \sum_{j\in\mathcal{H}} x_{ij}^{p,t} = 0, \quad i \in \mathcal{H}, t = 1, 2, \ldots, T, \tag{52}$$

$$x_i^{h,t} + \delta_i^t - x_i^{h,t-1} - \sigma_i^t - x_i^{h,t} = 0, \quad i \in \mathcal{H}, t = 1, 2, \ldots, T, \tag{53}$$

$$\sum_{j\in\mathcal{H}} x_{ij}^{c,t} - x_i^{l,t} - \sum_{j\in\mathcal{H}} x_{ij}^{p,t} = 0, \quad i \in \mathcal{H}, t = 1, 2, \ldots, T, \tag{54}$$

$$x_{ij}^{c,t} \leqslant \kappa_{ij}^t, \quad i \in \mathcal{H}, t = 1, 2, \ldots, T, \tag{55}$$

$$x_i^{r,t} = \delta_i^t - x_i^{d,t}, \quad i \in \mathcal{H}, t = 1, 2, \ldots, T. \tag{56}$$

Relations (55) specify that demand must be satisfied. At each port and period, Equations (52) compute the volume of containers unloaded from ships (the difference between the total volumes arriving and departing on all ships), relations (53) represent the inventory equations that also enforce the requirement that demand must be satisfied, and Equations (54) compute the number of containers being repositioned. Constraints (55) enforce the residual capacity for repositioning movements and Equations (56) define the number of leased containers as the difference between the number of owed containers available and demand. The formulation is represented by a network problem with random arc capacities and is cast as a two-stage stochastic program. Stochastic quasi-gradient and hybrid approximation solution procedures are proposed and are compared through a rolling-horizon experiment.

Powell and Carvalho (1998) address fleet management problems within the context of rail intermodal systems. The goal is to consider simultaneously the management of the fleet of flat cars the railway uses to provide service, the fleet of trailers and containers (collectively know as boxes) the railway owns and rents to its customers, and the complex rules that govern the substitutions among box types and the assignment of boxes to flat cars. The problem is highly dynamic and characterized by two sources of stochasticity: forecast of

customer demand and the return of boxes and flat cars from customers and other railways.

The authors decompose the problem into a problem addressing the management of the boxes the railway owns and a flat car fleet management problem. Each is formulated as a mixed-integer program over a multicommodity space–time diagram representing "all" possible movements of vehicles and loads (including holding activities) over the planning horizon. Decision variables concern departure times, the type of vehicle (and departure time) for each load, repositioning movements, flows of vehicles and loads. The objective maximizes the expected total profit of the system. Each formulation is then cast as a recursive dynamic model and, using approximations of the future values of vehicles and boxes at the nodes of the space–time network, it is decomposed into "easy-to-solve" local subproblems (assignment of boxes to requests or combinations of boxes to flat cars). In a series of forward–backward passes through the network, the algorithm then refines these approximations and assignments. Experimentations based on actual data suggest significant improvements over the planning procedures used in industry. More details on this methodology may be found in Powell (2003), Powell and Topaloglu (2003, 2005), and Powell et al. (2007).

5 Models for seaport container terminal operations

This section describes issues and introduces corresponding models for operational planning and control in port container terminals. The following subsections introduce models for berth scheduling, quay-crane scheduling, stowage planning and sequencing, storage space planning, and dispatching of yard cranes and transporters: yard trucks, straddle carriers and, for automated terminals, automated guided vehicles (AGV). Reviews can be found in Steenken et al. (2004) and in the book edited by Günther and Kim (2005).

5.1 Berth scheduling

As already mentioned, berths are the most important resource in port container terminals because berth construction costs are the highest among all relevant cost factors. *Berth scheduling* is the process of determining the time and position at which each arriving vessel will berth. *Quay-crane allocation* is the process of determining the vessel that each quay crane will serve and the time during which the quay crane will serve the assigned vessel. (The terms *scheduling* and *allocation* are often used interchangeably for both problems.) The berth scheduling and the quay-crane allocation problems are related because the number of quay cranes assigned to a vessel impacts the berthing duration of the ship. Despite this relationship, most studies treat the two issues separately to avoid the complexity of the integrated problem. The study by Park and Kim (2003) is an exception.

Ports have long used priority rules to determine the allocation of berths to incoming ships and the earliest studies focused on this approach (e.g., van der Heyden and Ottjes, 1985). Simulation was used to evaluate and compare rules (e.g., Lai and Shih, 1992). Brown et al. (1994, 1997) proposed the first mathematical models for allocating berths to vessels. Their studies focused on military vessels and assumed that each vessel required different services (re-provisioning, maintenance, repair, training, and certification test). Since not all services were provided at all berths, scheduling vessel shifts between berths was an important issue. This is different from the case of container ships for which only loading and unloading operations need to be considered.

Many ports are configured such that berths may be considered as subsections of a continuous line that ships of finite lengths can share, and several studies (Lim, 1998; Park and Kim, 2002; Park and Kim, 2003; Kim and Moon, 2003; Guan and Cheung, 2004) consider the berth allocation problem as a type of continuous line partitioning problem. Most researchers have treated berth scheduling as a discrete resource allocation problem, however. This approximation reduces the problem to that of assigning berths to arriving ships, which is much easier to address than the continuous berth-scheduling problem. We start the presentation with this second approach.

Cordeau et al. (2005) proposed an integer programming model for the discrete version of the berth-allocation problem based on a Multidepot Vehicle Routing Problem with Time Windows (MDVRPTW) formulation, where ships are seen as customers and berths as depots. One vehicle is located at each depot. Each vehicle starts and ends its tour at its depot. Ships are modeled as vertices in a multigraph. Every depot is divided into an *origin* vertex and a *destination* vertex. Time windows can be imposed on every vertex to represent the availability period of the corresponding berth. Notation is as follows:

N: Set of ships; $n = |N|$;
M: Set of berths; $m = |M|$;
t_i^k: Handling time of ship i at berth k;
a_i: Arrival time of ship i;
s^k: Start of availability time of berth k;
e^k: End of availability time of berth k;
b_i: Upper bound of the service time window of ship i;
v_i: Value of the service time for ship i;
$o(k)$: Starting operation time of berth k;
$d(k)$: Ending operation time of berth k;
$M_{ij}^k = \max\{b_i + t_i^k - a_j, 0\}$, $k \in M$, i and $j \in N$;
$G^k = (V^k, A^k)$, $k \in M$, where $V^k = N \cup \{o(k), d(k)\}$ and $A^k \subseteq V^k \times V^k$.

With decision variables:

x_{ij}^k: $x_{ij}^k = 1$, if and only if ship j is scheduled after ship i at berth k; 0, otherwise;

T_i^k: Berthing time of ship i at berth k, i.e., the time when the ship moors;
$T_{o(k)}^k$: Starting operation time at berth k, i.e., the time when the first ship moors at the berth;
$T_{d(k)}^k$: Ending operation time at berth k, i.e., the time when the last ship departs from the berth;

the MDVRPTW model may be written as:

$$\text{minimize} \quad \sum_{i\in N}\sum_{k\in M} v_i\left[T_i^k - a_i + t_i^k \sum_{j\in N\cup\{d(k)\}} x_{ij}^k\right] \tag{57}$$

subject to

$$\sum_{k\in M}\sum_{j\in N\cup\{d(k)\}} x_{ij}^k = 1, \quad i \in N, \tag{58}$$

$$\sum_{j\in N\cup\{d(k)\}} x_{o(k),j}^k = 1, \quad k \in M, \tag{59}$$

$$\sum_{j\in N\cup\{o(k)\}} x_{i,d(k)}^k = 1, \quad k \in M, \tag{60}$$

$$\sum_{j\in N\cup\{d(k)\}} x_{ij}^k - \sum_{j\in N\cup\{o(k)\}} x_{ji}^k = 0, \quad k \in M, i \in N, \tag{61}$$

$$T_i^k + t_i^k - T_j^k \leqslant (1 - x_{ij}^k)M_{ij}^k, \quad k \in M, (i,j) \in A^k, \tag{62}$$

$$a_i \leqslant T_i^k, \quad k \in M, i \in N, \tag{63}$$

$$T_i^k + t_i^k \sum_{j\in N\cup\{d(k)\}} x_{ij}^k \leqslant b_i, \quad k \in M, i \in N, \tag{64}$$

$$s^k \leqslant T_{o(k)}^k, \quad k \in M, \tag{65}$$

$$T_{d(k)}^k \leqslant e^k, \quad k \in M, \tag{66}$$

$$x_{ij}^k \in \{0,1\}, \quad k \in M, i,j \in A^k. \tag{67}$$

The objective is to minimize the weighted sum of the service times for all the vessels. Constraint (58) implies that each vessel must be assigned once to a berth. Constraints (59) and (60) define the degree of the depots. Constraint (61) enforces the flow conservation. The consistency of the T_i^k variables with the time sequence on a berth is guaranteed by constraints (62). Constraints (63) and (64) indicate the service time windows for vessels, while constraints (65) and (66) set the available time windows on the berths. For small instances, the above formulation can be solved by commercial software for integer linear programming models. For problem instances of realistic size, Cordeau et al. (2005) proposed a tabu search-based metaheuristic.

A different approach consists in explicitly representing and penalizing the difference between the berthing order implied by the ship priority and the

one proposed. The resulting models take the form of nonlinear integer programming formulations (Imai et al., 1997, 2001, 2003). Lagrangian-based and, for problem instances of realistic size, genetic metaheuristic solution methods have been proposed. The discrete berth-scheduling problem may also be cast as a machine-scheduling problem. Li et al. (1998) introduced a formulation for the scheduling of a single processor (the berth) that can simultaneously perform multiple jobs (vessels). The authors aimed to minimize the makespan and, based on the similarity of the problem to the bin-packing problem, suggested various algorithms based on the First-Fit-Decreasing heuristic. Guan et al. (2002) proposed an m-parallel machine scheduling formulation. In their model, the machines (quay cranes) are arranged along a straight line and each job (ship) requires simultaneous processing by multiple consecutive processors. Ships are characterized by size, processing time, and weight (priority). A heuristic was proposed to minimize the total weighted completion time of the jobs (ships).

A drawback of considering a berth as a collection of discrete berthing locations is that the number of ships that may be served simultaneously is fixed regardless of ship lengths. The continuous representation does not have this limitation: for the same length of berth, more vessels can be served simultaneously if they are shorter. On the other hand, in contrast to the discrete variant, the continuous berth-scheduling problem requires determining the exact berthing position of each ship as a real-valued position on a continuous line. Moreover, the berthing time for each ship must be determined simultaneously. The goals of the process include minimizing the ship departure delays and the container-handling costs that depend on the berthing position of each ship.

Lim (1998) was the first to consider a berth as a continuous line rather than a collection of discrete segments and viewed the berth planning problem as a two-dimensional bin packing problem. He discussed how to locate berthing positions of vessels so that the throughput of the berth is maximized, but did not consider the berthing time as a decision variable. Park and Kim (2002) introduced the first model for the continuous berth-scheduling problem to determine simultaneously the berthing time and position for each ship. The objective of the model is to minimize the costs resulting from delayed vessel departures, plus the additional handling costs resulting from deviations of the berthing position from the best location on the berth. The following information is assumed known:

L: Length of the berth;
l: Number of vessels;
a_i: Expected arrival time of vessel i;
b_i: The ship operation time required for vessel i;
d_i: Requested departure time of vessel i;
l_i: Length of vessel i; this value includes the required gap between adjacent vessels;

c_{1i}: Additional travel cost (per unit berth length) of delivering containers to vessel i resulting from nonoptimal berthing locations;
c_{2i}: Penalty cost (per time unit) of vessel i resulting from a delayed departure after the requested due time.

The problem is set in a two-dimensional space, the berth length and the planning time defining the horizontal and vertical axes, respectively. The reference point for the berth-length coordinate is the leftmost boundary of the berth. Figure 3 illustrates this setting, as well as the relationships between the input data and the following decision variables:

x_i: Berthing position of vessel i;
y_i: Berthing time of vessel i;
p_i: Best berthing location of vessel i; This location is represented by the x-coordinate (berth length axis) of the leftmost end of the vessel. It is determined by considering the distribution of containers to be loaded into the vessel;
z_{ij}^x: Equals 1, if vessel i is located to the left of vessel j in the time-berth-length space; 0, otherwise; Vessel i is located to the left of vessel j in Figure 3;
z_{ij}^y Equals 1, if vessel i is located below vessel j in the time-berth-length space; 0, otherwise; Vessel i is not considered to be located below vessel j in Figure 3 because of their partial overlap in the time dimension.

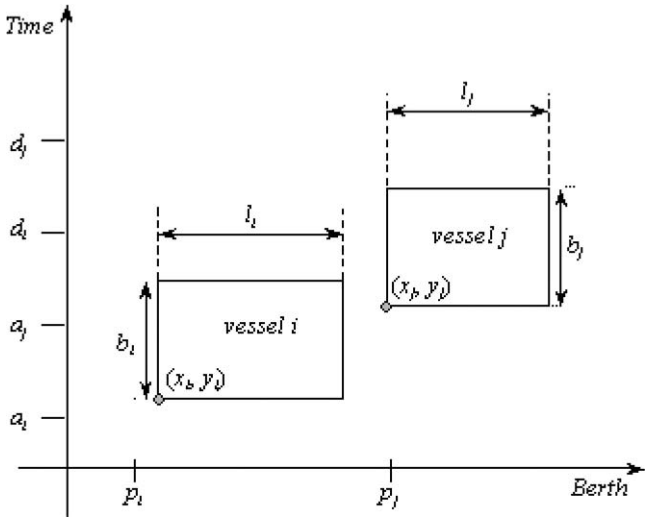


Fig. 3. Structure of continuous berth scheduling problem (Park and Kim, 2002).

An objective function can then be written as:

$$\text{minimize} \quad \sum_{i=1}^{l} \{c_{1i}|x_i - p_i| + c_{2i}(y_i + b_i - d_i)^+\}, \tag{68}$$

where $x^+ = \max\{0, x\}$. The first term of (68) computes the penalty for the deviation of the berthing positions from the best locations, while the second computes the penalty corresponding to the departure delays of the vessels leaving after the requested departure time. Let $|x_i - p_i|$ be α_i^+ when $x_i - p_i \geqslant 0$ and α_i^- when $x_i - p_i < 0$. Let $(y_i + b_i - d_i)$ be β_i^+ when $y_i + b_i - d_i \geqslant 0$ and β_i^-, otherwise. Then, the berth scheduling problem can be formulated as follows:

$$\text{minimize} \quad \sum_{i=1}^{l} \{c_{1i}(\alpha_i^+ + \alpha_i^-) + c_{2i}\beta_i^+\} \tag{69}$$

subject to

$$x_i - p_i = \alpha_i^+ - \alpha_i^-, \quad i = 1, 2, \ldots, l, \tag{70}$$

$$y_i + b_i - d_i = \beta_i^+ - \beta_i^-, \quad i = 1, 2, \ldots, l, \tag{71}$$

$$x_i + l_i \leqslant L, \quad i = 1, 2, \ldots, l, \tag{72}$$

$$x_i + l_i \leqslant x_j + M\big(1 - z_{ij}^x\big), \quad j = 1, 2, \ldots, l, i \neq j, \tag{73}$$

$$y_i + b_i \leqslant y_j + M\big(1 - z_{ij}^y\big), \quad i, j = 1, 2, \ldots, l, i \neq j, \tag{74}$$

$$z_{ij}^x + z_{ji}^x + z_{ij}^y + z_{ji}^y \geqslant 1, \quad i, j = 1, 2, \ldots, l, i < j, \tag{75}$$

$$y_i \geqslant a_i, \quad i = 1, 2, \ldots, l, \tag{76}$$

$$\alpha_i^+, \alpha_i^-, \beta_i^+, \beta_i^-, x_i \geqslant 0, \quad i = 1, 2, \ldots, l, \tag{77}$$

$$z_{ij}^x,\ z_{ij}^y \in \{0, 1\}, \quad i, j = 1, 2, \ldots, l,\ i \neq j. \tag{78}$$

Constraints (70) and (71) define α_i^+, α_i^-, β_i^+, and β_i^-. Constraints (72) specify that the position of the rightmost end of vessel i is restricted by the length of the berth. A constraint of the type (73) or (74) is effective only when the corresponding z_{ij}^x or z_{ij}^y, respectively, equals one. The constraints define the relative "left" and "bellow" position of vessel i relative to that of vessel j in the time-berth-length space. Constraints (75) enforce the condition that two vessels cannot occupy the same space during the same time. The constraints proceed by excluding in the time-berth-length diagram the case $z_{ij}^x + z_{ji}^x + z_{ij}^y + z_{ji}^y = 0$, which corresponds to schedule overlapping for vessels i and j. Constraints (76) imply that vessels cannot berth before they arrive.

Model (69)–(78) is a mixed-integer programming formulation. Park and Kim proposed a subgradient optimization technique and solved problems with 13–20 ships in a few minutes of computational time on a Pentium II (233 MHz and 128 Mb RAM) computer. Similar formulations have also been proposed

by Kim and Moon (2003) and Guan and Cheung (2004). Kim and Moon examined a stability property that berthing locations of vessels must satisfy to form an optimal solution and used it in a simulated annealing procedure. Guan and Cheung used a different objective function. They minimized the total weighted flow time instead of the penalties for delayed service and the deviation of berthing positions from the best locations of Park and Kim (2002). Guan and Cheung proposed a tabu search metaheuristic based on pair-wise exchanges, which appears to perform well.

5.2 *Quay-crane scheduling*

Quay cranes are another important resource in container terminals. In practice, the quay-crane schedule is usually constructed by operation planners as part of ship plans. Two types of quay-crane scheduling problems have been defined according to the detail of the representation. The first problem simultaneously schedules the berth and the quay cranes by specifying the starting and ending times of unloading and loading operations for each quay crane assigned to a specific ship. Park and Kim (2003) proposed a model and a two-phase solution procedure. The first phase determines the berthing position and time of each vessel as well as the number of quay cranes assigned to each vessel at each time segment. Subgradient optimization is used to obtain a near-optimal solution of the first phase. In the second phase, a detailed schedule for each quay crane is constructed by dynamic programming starting from the solution of the first phase.

The second type of quay-crane scheduling problem determines the detailed schedule for each quay crane assigned to a vessel for a pre-specified time window to carry out container unloading or loading operations. Most studies on quay-crane scheduling address this problem. The modeling methodology generally used is based on integer programming models. Daganzo (1989b) addressed the quay-crane scheduling problem in port terminals for general and containerized cargo. He attempted to determine the number of quay cranes to assign to each ship bay at each time segment. An integer programming model was suggested with an objective to minimize the completion time of ship operations for all the ships.

Peterkofsky and Daganzo (1990) refined the previous model and proposed a branch-and-bound algorithm. In their model, S ships are moored alongside the berth of a terminal. Ship i has H_i holds numbered $(i, 1), (i, 2), \ldots, (i, H_i)$. The time required to complete the loading and unloading operations for hold (i, j) is denoted W_{ij} measured in quay-crane-hours. It is assumed that all ships are ready for loading/unloading operations at time $t = 0$. It is also assumed that sufficient berth space is provided for all the ships to berth at the same time. M quay cranes are available. Decision variables are $y_{ijm}(t) = 1$ if quay crane m, $m = 1, 2, \ldots, M$, is allocated to hold (i, j) at time t, and 0 otherwise. The objective function minimizes the weighted amount of time that ships spend in port, $\sum_{i=1}^{S} C_i T_i$, where T_i and C_i stand for the departure time and

cost per hour of delay for ship i, respectively. A ship can depart only when the operations on all its holds have been completed, that is, if

$$\sum_{m=1}^{M} \int_0^{T_i} y_{ijm}(t)\,\mathrm{d}t \geqslant W_{ij}, \quad j = 1, 2, \ldots, H_i. \tag{79}$$

In some cases, there may exist other activities than ship unloading and loading which must be performed (e.g., refill of supplies or repairs) and thus the earliest possible departure time may be restricted by a limit τ_i yielding the constraints

$$T_i \geqslant \tau_i \quad \text{for } i = 1, 2, \ldots, S. \tag{80}$$

A quay crane may be allocated to no more than one hold at a time. Thus, for any value of t, the crane assignments must satisfy:

$$\sum_{i=1}^{S} \sum_{j=1}^{H_i} y_{ijm}(t) \leqslant 1, \quad m = 1, 2, \ldots, M. \tag{81}$$

Finally, when only a maximum of Δ_{ij} quay cranes can be allocated to hold (i, j), the following constraints must be added to the formulation:

$$\sum_{m=1}^{M} y_{ijm}(t) \leqslant \Delta_{ij}, \quad i = 1, 2, \ldots, S,\ j = 1, 2, \ldots, H_i. \tag{82}$$

Peterkofsky and Daganzo (1990) assumed a situation with no serious interference between cranes and, consequently, did not impose any restrictions on the movements of cranes. This approach is appropriate for general cargo handling. However, quay cranes in container terminals travel on the same rail, which results in various interference possibilities between adjacent cranes. Consequently, adjacent quay cranes need space of at least two ship bays between them. At the same time, no crane can pass over adjacent cranes, which makes the problem more complicated. Kim and Park (2004) addressed the quay-crane scheduling problem at this level of detail, but for a single vessel only. The model determines starting and ending times for each quay crane to serve each ship bay, under various constraints representing detailed movements of quay cranes and interferences among quay cranes. The authors proposed both a branch-and-bound algorithm and a GRASP-based heuristic to overcome the computational difficulty of solving the mixed-integer programming formulation exactly.

Comparing the work of Peterkofsky and Daganzo (1990) and Kim and Park (2004), one realizes that the difference comes from the different viewpoints on how the whole ship planning problem should be addressed. Peterkofsky and Daganzo (1990) considered the quay-crane scheduling as a part of berth planning, whereas Kim and Park (2004) separated the detailed crane-scheduling problem from the berth-scheduling problem. This comparison also provides

an illustration of the often-encountered trade off between the integration of several planning issues into the same model and the level of detail one can include in the representation of each element of the system being considered.

5.3 Stowage planning and sequencing

A *container group* is defined as a collection of containers of the same size and with the same destination port. *Stowage planning* determines the block (cluster of adjacent slots) of a ship bay a specific group of containers should be stacked into. The stowage planning process must consider the burden of additional container manipulations when containers bound for succeeding ports are located in higher tiers than those that must be unloaded at a given port. The stowage plan must also comply with various measures of ship stability and strength. Christiansen et al. (2007) review studies on stowage planning.

The *stowage-sequencing* problem must be addressed next by determining the sequence of unloading inbound containers and loading outbound containers. The slot into which each of the outbound containers will be stacked must be determined simultaneously. When an indirect transfer system is used, yard cranes move containers between yard stacks and yard trucks. The loading sequence of individual containers thus impacts significantly the total distance traveled by yard cranes and hence the total handling cost in the yard. This is not the case when a direct transfer system is used since the yard handling cost is determined by the cost of yard truck movements only, which is independent of the container loading sequence.

In the stowage (load/unload) sequencing problem, it is usually assumed that the stowage plan is already constructed and provided by the vessel carrier. The problem is then reduced to assigning slots to individual loading containers and sequencing unloading and loading operations. Most research has focused on the sequence of loading operations, which influences the handling costs more significantly than the sequence of unloading operations. Indeed, containers to be loaded into the slots of a ship must satisfy various constraints on the slots, which are pre-specified by the stowage planner. Furthermore, locations of outbound containers may be scattered over a wide area in a terminal yard. The time required for loading operations, thus, depends on the cycle time of both quay and yard cranes. The cycle time of a quay crane depends on the loading sequence of slots in the vessel, while the cycle time of a yard crane is affected by the loading sequence of containers in the yard.

Research on load sequencing can be classified according to the scope of the problem. Several efforts have addressed the *pickup-scheduling* problem in which the travel route of each yard crane and the number of containers to be picked up at each yard bay on the route are determined (Kim and Kim, 1999b, 1999c; Kim and Kim, 1999d, 2003; Narasimhan and Palekar, 2002; Ryu et al., 2001). Other authors have attempted to determine the loading sequence of individual containers present in a yard into the slots of a vessel, a process that requires more detailed scheduling than the previous problem (Cojeen and

Table 2.
Yard map

Yard-bay	1	2	4	5	7	8	9	11	12	14	15
Container group	A	B	C	B	A	B	C	B	A	C	A
Number of containers	14	14	13	14	12	7	12	8	13	10	13

Table 3.
Load plan

Subsequence number	1	2	3	4	5	6	7	8
Container group	A	B	A	C	B	A	B	C
Number of containers	15	20	24	25	11	23	12	10

Dyke, 1976; Kozan and Preston, 1999; Kim et al., 2004). In this subsection, we focus on the pickup-scheduling problem, with only a brief literature review dedicated to the individual-container-scheduling problem.

The pickup-scheduling problem for a piece of yard equipment (e.g., a yard crane) can be described as follows (Kim and Kim, 1999b; Narasimhan and Palekar, 2002): In a terminal yard, containers are classified into groups, each yard bay holding a number of containers of a number of particular groups. Table 2 illustrates a yard map, that is, the distribution of containers by container group over the yard bays. It is a simplified case where each yard bay contains containers of one group only: 14 containers of group A are in yard-bay 1, 14 containers of group B are in yard-bay 2, and so on. A load plan specifies a sequence of blocks of pickup operations, each block representing a subsequence of consecutive pickups of containers belonging to the same group. Table 3 shows a load plan composed of eight subsequences to pick up sequentially 15 containers of group A, 20 containers of group B, 24 of group A, and so on. A partial tour denotes a sequence of yard bays the yard crane visits to perform the operations for a given subsequence. The pickup-scheduling problem is to decide (1) the container locations (yard bays) to use for each subsequence; and (2) the partial tour for each subsequence, i.e., the visiting order of the yard bays assigned to each subsequence in (1). The objective is to minimize the total setup and travel time the yard crane requires to pick up all the containers in the load plan.

The following pickup-scheduling model follows Kim and Kim (1999b). It addresses the scheduling of a single yard crane. The problem parameters are:

m: Number of subsequences (and partial tours) that constitute a complete tour for a yard crane;
n: Number of yard bays;
l: Number of container groups;
c_{hj}: Initial number of containers of group h stacked at yard bay j;
t: Subsequence number, $t = 1, 2, \ldots, m$;

r_t: Number of containers in subsequence t;
g_t: Container-group number of subsequence t;
$S(h)$: Set of subsequences for the container-group number h;
$B(h)$: Set of the yard bays where containers of group h are stored;
S: Initial location of the yard crane, noted $B(g_0) = \{S\}$;
T: Final location of the yard crane, noted $B(g_{m+1}) = \{T\}$;
d_{ij}: Travel distance of the yard crane between yard bays i and j;
e_{ij}: Equals 1, if yard bays i and j are distinct and 0, otherwise (i.e., $i = j$);
T_s: Setup time of the yard crane for each visit to a yard-bay;
T_d: Travel time of the yard crane per the distance of a yard-bay length.

The decision variables are:

$y_{ij}^t = 1$, if the yard crane moves from yard bay i to yard bay j after completing the partial tour t; 0, otherwise; y_{Sj}^0 and y_{jT}^m denote the first and the final movements of the yard crane during the tour, respectively;

$z_{ij}^t = 1$, if the yard crane moves from yard bay i to yard bay j during the partial tour t; 0, otherwise;

$x_j^t =$ Number of containers picked up at yard bay j during the partial tour t.

The problem may then be formulated as follows:

$$\begin{aligned}\text{minimize}\quad & T_s\left(\sum_{t=0}^{m}\sum_{i\in B(g_t),\,j\in B(g_{t+1})} e_{ij}y_{ij}^t + \sum_{t=1}^{m}\sum_{i,j\in B(g_t)} z_{ij}^t\right)\\ & + T_d\left(\sum_{t=0}^{m}\sum_{i\in B(g_t),\,j\in B(g_{t+1})} d_{ij}y_{ij}^t + \sum_{t=1}^{m}\sum_{i,j\in B(g_t)} d_{ij}z_{ij}^t\right)\\ & = \sum_{t=0}^{m}\sum_{i\in B(g_t),\,j\in B(g_{t+1})} (T_s e_{ij} + T_d d_{ij})y_{ij}^t\\ & \quad + \sum_{t=1}^{m}\sum_{i,j\in B(g_t)} (T_s e_{ij} + T_d d_{ij})z_{ij}^t \qquad (83)\end{aligned}$$

subject to

$$\sum_{j\in B(g_1)} y_{Sj}^0 = 1, \qquad (84)$$

$$-\sum_{j\in B(g_m)} y_{jT}^m = -1, \qquad (85)$$

$$\sum_{j\in B(g_{t-1})} y_{ji}^{t-1} - \sum_{j\in B(g_{t+1})} y_{ij}^{t} + \sum_{k\in B(g_t)} \left(z_{ki}^{t} - z_{ik}^{t}\right) = 0,$$
$$i \in B(g_t),\ t = 1, 2, \ldots, m, \tag{86}$$

$$\sum_{i,j\in B(g_t)} z_{ij}^{t} \leqslant |N| - 1, \quad N \subseteq B(g_t),\ t = 1, 2, \ldots, m, \tag{87}$$

$$x_j^t \leqslant M\left(\sum_{k\in B(g_t)} z_{kj}^{t} + \sum_{i\in B(g_{t-1})} y_{ij}^{t-1}\right),$$
$$j \in B(g_t),\ t = 1, 2, \ldots, m, \tag{88}$$

$$\sum_{j\in B(g_t)} x_j^t = r_t, \quad t = 1, 2, \ldots, m, \tag{89}$$

$$\sum_{t\in S(h)} x_j^t \leqslant c_{hj}, \quad h = 1, 2, \ldots, l,\ j \in B(h), \tag{90}$$

$$y_{ij}^{t} \in \{0, 1\}, \quad i \in B(g_t),\ j \in B(g_{t+1}),\ t = 1, 2, \ldots, m, \tag{91}$$

$$z_{ij}^{t} \in \{0, 1\}, \quad i, j \in B(g_t),\ t = 1, 2, \ldots, m, \tag{92}$$

$$x_j^t \geqslant 0, \quad j \in B(g_t),\ t = 1, 2, \ldots, m, \tag{93}$$

where M is a sufficiently large number and $|N|$ denotes the cardinality of set N.

The objective function (83) minimizes the total time required by the yard crane to perform the load plan, which depends on the total number of setups and the total distance traveled. Because a setup occurs whenever a yard crane moves from one yard bay to another, only the inter-yard-bay movements are considered in the evaluation of the total number of setups. A feasible solution corresponds to a path from the source node to the terminal node of the network. Constraints (84) and (85) represent conservation of flow at the source and terminal nodes, respectively, while constraints (86) enforce the flow conservation at the other nodes. Relations (87) are subtour elimination constraints that exclude isolated cycles (i.e., not connected to a path from the source to the destination node) from a partial tour. Constraints (88) imply that only when a yard crane visits a yard bay can it pick up containers there. Constraints (89) enforce the condition that the number of containers picked up in a partial tour must equal the number of containers requested by the load plan. Constraints (90) indicate that the total number of containers picked up during the entire tour must not be larger than the initial number of containers available at each yard bay.

To build this formulation, Kim and Kim (1999b) took advantage of the limited motion of yard cranes, which move on a straight rail line, to reduce the solution space. Kim and Kim (1999d) and Kim and Kim (1999c) generalized this model for the pickup-scheduling problem for a single straddle carrier and for multiple straddle carriers, respectively. Kim and Kim (2003) compared exact optimization, a beam search heuristic, and a genetic metaheuristic for solving the model proposed by Kim and Kim (1999b). Based on objective-function

value, the beam search outperformed the genetic metaheuristic on problems with 30 yard bays for which exact optimization algorithms could not be used due to excessive computing times. Ryu et al. (2001) suggested an ant system metaheuristic for the same formulation and compared it to a tabu search metaheuristic on problem instances with 39–514 ships and 36–350 yard bays. On average, the ant colony metaheuristic was 49 times slower than the tabu search but achieved a solution value 17% better.

Narasimhan and Palekar (2002) analyzed a generalized version of the scheduling problem proposed by Kim and Kim (1999b). The authors relaxed the assumption that the order of subsequences is fixed, which may yield a reduction in the total handling time of transfer cranes. They also proved a number of properties of the optimal solution to this version of the problem, in particular that:

(1) the yard bays visited by a transfer crane during a partial tour must be contiguous in the yard-bay map;

(2) sets of yard bays for partial tours of the same container group coincide with partitions of the yard-bay map; and

(3) the minimization of the total distance traveled by transfer cranes results in the minimization of the transfer-crane setup time.

The authors then restated the problem as an odd-node-matching problem where edges are added at minimum cost such that every node of the graph has en even degree. They proved that this version of the problem is NP-hard in the strong sense and proposed both a branch-and-bound and an enumerative heuristic to address it.

The problem of constructing a containership loading plan at the level of individual containers appears more complex than the pickup-scheduling one. On the one hand, the problem is larger because the loading sequence of individual containers must be determined. On the other hand, the goals and limitations related to the operation of various pieces of equipment, quay and transfer cranes notably, and by the constraints imposed by the stowage plan must be accounted for (Kim et al., 2004). The objectives related to the operation of quay cranes are first to fill slots in the same hold, then stack containers onto the same tier on deck, and finally stack containers of weights included in the same weight group as specified in the stowage plan. The objectives related to the operation of transfer cranes are to minimize the travel time of transfer cranes, minimize the number of rehandles (the same container that is moved more than once to, for example, allow access to containers stacked underneath), and pick up containers in locations nearer to the transfer point earlier than those located farther from the transfer point. The constraints related to the operation of quay cranes are to follow precedence relationships among slots that follow from their own work schedules and the relative positions of slots in a ship bay, not violate the maximum allowed total weight of the stack on deck, not violate the maximum allowed height of the stack of a hold, and load the same type of containers as specified in the stowage plan. A constraint related

to the operation of transfer cranes is to maintain sufficient distance between adjacent transfer cranes to avoid interferences.

The early contributions presented software based on heuristic rules and simple procedures (e.g., Cojeen and Dyke, 1976). Kim et al. (2004) proposed an integer linear programming model for sequencing individual outbound containers. The cost-minimization objective accounted for the handling cost of both quay and transfer cranes, while constraints included restrictions on the maximum weight allowed on deck, maximum height of stacks in holds, and the stability of the vessels. The authors proposed a beam search heuristic for this formulation. Kozan and Preston (1999) focused on the scheduling of transfer operations where the actual sequencing issues are not considered. The authors proposed a mixed-integer formulation to determine the storage locations of containers in the yard, the yard equipment to move the containers, and the schedules of these movements, such that the total travel cost of the yard equipment is minimized. A genetic metaheuristic was proposed as solution method.

5.4 Storage activities in the yard

The scope of most research dedicated to the operation of yards is to minimize the container handling efforts and utilize the storage space efficiently. In addressing these issues, one must consider the rather different characteristics of inbound (usually import) and outbound (export) container flows. Inbound containers are usually unloaded fast not to delay the ship departure, but are retrieved over a long period of time compared to the ship unloading time, in a somewhat random sequence due to uncertainties related to importing formalities and the operations of land transportation modes. The minimization of the number of container rehandlings is, thus, the most important issue in this case. The same characteristics of the land and maritime transportation modes make outbound containers arrive at the yard randomly over a long period of time, while their loading is performed rapidly. Pre-planning storage spaces for arriving containers will result in less rehandlings and a more efficient ship loading operation.

A number of early studies contributed greatly to the understanding of these problems. They described practical issues related to space allocation, analyzed basic properties related to container handling in storage yards, and proposed formulas to evaluate yard productivity measures such as the total handling effort or the changes in container inventory due to various container-handling strategies (Taleb-Ibrahimi et al., 1993; Castilho and Daganzo, 1993; Kim, 1997; Chen, 1999).

The *space-allocation* problem is concerned with determining storage locations for containers either individually or as a group. The problem may focus on inbound or outbound container movements, or it may consider both. Storage activities and requirements in yards may be initiated by several sources, such as export containers arriving through gates by truck, at quay by feeder vessels, or by rail service, as well as import containers unloaded from vessels

and moved out of the terminal by trucks, feeder vessels, or rail. Each storage activity is characterized by the amount of storage space required, the source and destination of the containers involved (e.g., export containers unloaded from rail have rail as source and the berth as destination), and the starting and ending times of the storage activity. Information regarding these factors can be obtained from delivery schedules or forecasts, as well as unloading/loading schedules for vessels.

We present a formulation that follows that of Kim and Park (2003b) and may be applied to the planning of storage locations for both inbound and outbound containers. The formulation is defined for a planning horizon discretized in T periods, $t = 1, 2, \ldots, T$, according to the starting and ending times of storage activities. Let a_i and b_i represent the starting and the ending times of storage activity $i = 1, 2, \ldots, l$, respectively, and d_i the associated storage space required (i.e., the number of containers in twenty-foot equivalent units). The space-allocation problem may then be expressed as a minimum cost multicommodity network flow problem on a time–space network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$. Each storage activity i generates a *source* node S^i and a *destination* node T^i. *Intermediate* nodes j_t identify storage locations $j = 1, 2, \ldots, n$ at time periods t. Arcs from each source node S^i to intermediate nodes in period a_i represent the possible movements of containers from their origin toward the storage locations at the starting time of storage activity i. Symmetrically, arcs from intermediate nodes to a destination node T^i in period b_i stand for the possible movements of containers from the storage locations to their destination at the ending time of storage activity i. The corresponding transportation costs are associated to each of these two types of arcs, but no limits are imposed on their capacity. Holding arcs are defined between intermediate nodes representing the same storage location at consecutive time periods. The capacity u_j of holding arc (j_t, j_{t+1}) corresponds to the stowage capacity of the associated stacking area $j = 1, 2, \ldots, n$, for $t = 1, 2, \ldots, T$. Holding costs are not included in the formulation.

A storage activity i can be described as a route R_{ij} in $\mathcal{G}$ from node S^i to one of the intermediate nodes (storage location), through several holding arcs, to node T^i. The number of such routes equals the number of available storage locations:

$$R_{ij} : S^i \to j_{a_i} \to j_{a_i+1} \to j_{a_i+2} \to \cdots \to j_{b_i-1} \to j_{b_i} \to T^i.$$

A unit transportation cost c_{ij} is associated to each route R_{ij} of storage activity i. Let $\delta_{R_{ij}}^{j_t}$ equal 1 if arc (j_t, j_{t+1}) is included in route R_{ij} of storage activity i and 0, otherwise.

Define the decision variables x_{ij} as the number of containers of storage activity i that moves following route R_{ij}. The space-allocation problem can then be formulated as follows:

$$\text{minimize} \quad \sum_{i=1}^{l} \sum_{k=1}^{n} c_{ij} f_{ik} \tag{94}$$

subject to

$$\sum_{i=1}^{l} \delta_{R_{ij}}^{j_t} f_{ik} \leqslant u_j, \quad j = 1, 2, \ldots, n, t = 1, 2, \ldots, T, \tag{95}$$

$$\sum_{j=1}^{n} f_{ik} = d_i, \quad i = 1, 2, \ldots, l, \tag{96}$$

$$f_{ik} \geqslant 0 \text{ and integer}, \quad i = 1, 2, \ldots, l, j = 1, 2, \ldots, n. \tag{97}$$

The objective function (94) minimizes the total transportation cost of all the storage activities. Constraints (95) enforce the storage space limitations at each storage location and time period. Constraints (96) indicate that the space requirement of each storage activity must be satisfied. Notice that storage activities have been defined with constant space requirements for all time periods. A storage activity with a time-varying space requirement may be represented in this formulation as a combination of multiple storage activities with different starting and ending times and space requirement, respectively.

Formulation (94)–(97) corresponds to an integer, capacitated, minimum cost multicommodity flow problem, which is NP-hard (assuming all storage activities begin and end at the same time, the problem reduces to a knapsack problem, which is a well-known NP-hard problem). Moreover, real-world problems are frequently of very large dimensions. Thus, for example, the formulation for a container terminal with 20 blocks and 4 berths yields a mathematical model with more than 100,000 storage activities. Most contributions that focused on the container terminal storage activities thus propose heuristic solution methods. Methodology based on mathematical programming methods for large-scale integer multicommodity network flow problems has been proposed by Barnhart et al. (1998) and Barnhart et al. (2000).

Cao and Uebe (1995) addressed the static, single-period version of the previous problem and proposed a transportation problem formulation. Zhang et al. (2003) generalized the multiperiod problem of Kim and Park (2003b) by including into the objective function terms representing the balancing of handling activities among different storage areas, as well as the minimization of the total distance traveled by yard trucks between storage areas and quay cranes. The authors considered the flows of both inbound and outbound containers. They decomposed the problem into two stages. The first determined the total number of unloaded (from ships) and received (from land modes) containers to balance their distribution among yard blocks. The second stage determined the number of containers associated to each vessel allocated to each yard block in order to minimize the total travel cost of yard trucks. The first stage problem was formulated as a linear programming model, while the second took the form of a transportation problem.

Several other studies addressed aspects of the storage-space allocation problem different from Kim and Park (2003b). Kim and Bae (1998) examined transportation and dynamic programming models to schedule the remar-

shaling operations for outbound containers. Kim et al. (2000) addressed the problem of locating individual outbound containers considering the container weights. Kim and Park (2003a) focused on the space-allocation problem for export containers only. They considering various practical constraints and proposed integer programming formulations. Kim and Kim (1999a) addressed the space-allocation problem for import containers in which the stacking height and the amount of space (not the storage locations) are simultaneously determined to accommodate dynamically changing space requirements. The authors proposed a nonlinear programming model and a solution method based on Lagrangian relaxation. Kozan (2000), Preston and Kozan (2001), and Mattfeld and Kopfer (2004) proposed integer linear programming models to simultaneously determine storage locations and schedules of transfer operations. Castilho and Daganzo (1991) and Holguin-Veras and Jara-Diaz (1999) proposed methods for determining prices of storage spaces for containers or general cargo.

5.5 *Allocation and dispatching of yard cranes and transporters*

Yard equipment is an important container terminal resource that includes yard cranes and transporters, i.e., yard trucks and straddle carriers. In automated container terminals, automated guided vehicles belong to this category. The main operational decision issue at this level concerns the assignment of container-handling tasks to different pieces of equipment. Yard equipment is considered less critical than berths, storage spaces, and quay cranes for which schedules are constructed several hours, or even days, in advance. Consequently, no such planning is performed for yard equipment. Instead, decisions on allocation and dispatching yard equipment are made only for the limited number of activities to be carried out in the near future (e.g., one hour and less).

The objectives of allocating and dispatching yard cranes and transporters are usually to minimize the total travel distance, the total waiting time, or the total delays of equipment beyond prespecified due times. The dispatching problem for transporters was studied by Bish (2003), Kim and Bae (1999, 2004), Böse et al. (2000), and Grunow et al. (2004). Zhang et al. (2002), Cheung et al. (2002), Kim et al. (2003), Lai and Lam (1994), and Lai and Leung (1996) addressed the allocation and dispatching problem for yard cranes.

Consider two ships, sh^- and sh^+, that need unloading and loading, respectively, and that are berthed around the same time so that they can be served by the same set of k vehicles. It is assumed, however, that no ship-to-ship movement is possible. Let $\mathcal{N}^-$ and $\mathcal{N}^+$ denote the sets of containers that will be unloaded (from ship sh^-) and loaded (onto ship sh^+), respectively. Let c_i be container i in $\mathcal{N}^-$. Associated to each container to be loaded onto a ship is its *current storage location* in the yard area, which is known. Each such container will require a *loaded vehicle trip* from its current storage location to the location

of ship sh^+. Let L^+ denote the set of current storage locations of the containers in $\mathcal{N}^+$. Symmetrically, a set of *potential storage locations* in the yard area is reserved for the containers of each unloading ship. Set L^- contains the potential storage locations reserved for all containers to be unloaded from sh^-. Each unloaded container will require a *loaded vehicle trip* from the location of ship sh^- to its *selected* storage location. We make the simplifying assumption that sets L^+ and L^- are disjoint, that is, no container in $\mathcal{N}^-$ can be stored in a location currently occupied by a container in $\mathcal{N}^+$. Let $L = L^- \cup L^+$ and W_i be the subset of L^- where c_i can be stacked.

When the destination of a loaded trip differs from the origin of the next loaded trip on a vehicle schedule, the vehicle needs to make an empty, *repositioning* movement. The total amount of empty vehicle trips thus depends on the sequence of loaded trips of each vehicle and should be minimized by, for example, matching each loaded trip for an unloaded container to a loaded trip for a loading container. The goal therefore is to (1) determine a storage location for each unloaded container, (2) construct round vehicle trips, and (3) assign round trips to each vehicle. The first problem consists in assigning each unloaded container $c_i \in \mathcal{N}^-$ to a candidate storage location $p \in W_i$. Constructing round vehicle trips, the second problem, corresponds to matching a trip loaded with an unloaded container with a trip loaded with a loading container. Finally, the third problem consists in constructing sequences of round trips and assigning these to vehicles to minimize the make-span.

The following formulation follows the model proposed by Bish (2003) to simultaneously address the first two problems: determine storage locations for containers being unloaded from ships, schedule unloading and loading operations, and dispatch AGVs. The model is based on a network where supply nodes, with unit supply, correspond to unloaded containers $c_i \in \mathcal{N}^-$, demand nodes, with unit demands, correspond to current storage locations of loading containers l_q, $q \in L^+$, and transshipment nodes l_p stand for potential storage locations $p \in W_i$ reserved for unloading containers. The arc set is given by $\mathcal{A} = \{(c_i, l_p) \mid c_i \in \mathcal{N}^-, p \in W_i\} \cup \{(l_p, l_q) \mid p \in L^-, q \in L^+\}$, each arc with unit capacity. Arc $(c_i, l_p) \in \mathcal{A}$ corresponds to the trip of an unloading container c_i that is to be stored in location l_p; Its cost t_{lp} is thus the corresponding travel time. Arc $(l_p, l_q) \in \mathcal{A}$ corresponds to the empty trip of a vehicle that just completed the delivery on an unloaded container to location l_q. The vehicle moves to location l_q to pick up a loading container. The cost of the arc λ_{pq} is thus the travel time of the empty movement.

Define the decision variables x_{uv}, $(u, v) \in \mathcal{A}$, that equal 1 if the activity corresponding to arc (u, v) is to be performed (the arc is selected) and 0, otherwise. We can now model the problem as a transshipment formulation:

$$\text{minimize} \quad \sum_{c_i \in \mathcal{N}^-} \sum_{p \in W_i} t_{lp} x_{c_i l_p} + \sum_{p \in L^-} \sum_{q \in L^+} \lambda_{pq} x_{l_p l_q} \tag{98}$$

subject to

$$\sum_{p \in W_i} x_{c_i l_p} = 1, \quad c_i \in \mathcal{N}^-, \tag{99}$$

$$\sum_{c_i \in N^-} x_{c_i l_p} = \sum_{q \in L^+} x_{l_p l_q}, \quad p \in L^-, \tag{100}$$

$$\sum_{p \in L^-} x_{l_p l_q} = 1, \quad q \in L^+, \tag{101}$$

$$x_{uv} \in \{0, 1\} \quad \text{for all } (u, v) \in \mathcal{A}. \tag{102}$$

The objective function (98) minimizes the total assignment and matching-related travel time of the yard equipment. Constraints (99) ensure that each unloaded container is assigned to exactly one loaded trip and thus, to exactly one unloading location. Constrains (100) enforce the flow-balance requirements at storage sites for unloading containers. Relations (101) ensure that each loaded trip with an unloaded container is matched with a loaded trip with a loading container. The integrality conditions (102) imposed on the decision variables are not really required due to the total unimodularity property the formulation. The author proposed a heuristic for realistically-sized problem instances. The solution of the model yields a set of round trips for vehicles, which consist of a loaded movement with an unloaded container to a storage location, an empty repositioning trip from this location to a pickup location of a loading container, a loaded movement from this pickup location to a loading position under a quay crane, and an empty travel from there to the unloading position of another quay crane. These round trips are then assigned to vehicles by a list-scheduling heuristic.

Several other studies have been dedicated to issues related to yard equipment allocation and dispatching. Kim and Bae (1999) addressed the dispatching problem of AGVs in automated container terminals under the assumption that storage locations of containers as well as the schedules for unloading and loading operations by quay cranes are given. The scope is thus narrower than that of Bish (2003). Kim and Bae (1999) considered quay cranes are the most expensive equipment in container terminals. The model they proposed aimed therefore to minimize the total idle time of a quay crane resulting from late arrivals of AGVs as well as the associated total travel time. The authors suggested a network-based mixed-integer linear model and provided a heuristic algorithm. Nodes in the network represent "events" in time and space that correspond to moments (the event time) when vehicles pick up containers. An arc from one node to another indicates that the time lapse between the event times of the two operations allows a vehicle to drop off a container at the destination node after completing its pickup at the origin node. The problem is then to find the optimal routes on the network, each representing a sequence of delivery tasks assigned to a specific vehicle. Kim and Bae (2004) extended this approach to the case with multiple quay cranes in which dual cycle operation of AGVs is allowed.

Böse et al. (2000) addressed the dispatching problem of straddle carriers with the objective of minimizing the delays of quay cranes. They also assumed that storage locations of containers are already determined. The problem then becomes to assign delivery tasks of unloading and loading containers to straddle carriers. The problem is basically the same as the one addressed by Kim and Bae (2004) except that Böse et al. (2000) do not consider the vehicle travel times. The authors did not provide a formal model but proposed a genetic metaheuristic to search for solutions. Bish et al. (2005) represented the dispatching of vehicles as a machine-scheduling problem and proposed a greedy heuristic that assigned deliveries (jobs) to the first available vehicle (machine) which may arrive on time at the designated destination location, i.e., the quay crane or the location of the loading container. The authors showed that the greedy heuristic yields a job sequence that minimizes the makespan. Grunow et al. (2004) generalized the previous studies by considering the dispatching problem for multiload vehicles in automated container terminals. Indeed, while previous studies assumed that a vehicle can move only one container at a time, the authors considered the fact that a vehicle which moves one forty-foot container can move two twenty-foot containers. They proposed a mixed-integer linear programming model and suggested priority rules for dispatching delivery tasks to two-load AGVs.

Similar studies have been done for yard cranes. The objectives considered in most cases were the total waiting time and the total delays of trucks. Zhang et al. (2002) and Cheung et al. (2002) solved static versions of the crane deployment problem when the total workload at each storage area is known in advance. Zhang et al. (2002) proposed a mixed-integer programming model and addressed it by a method based on Lagrangian relaxation. Cheung et al. (2002) addressed a similar problem but removed the restriction that crane movements must be completed within a single period. This allows to use a shorter period length resulting in a more accurate model. A successive piecewise-linear approximation method was proposed. Kim et al. (2003) addressed the problem of sequencing transfer tasks of a yard crane for outside trucks in dynamic situations where new trucks arrive continuously. A dynamic programming model was suggested and decision rules derived by a reinforcement learning technique were proposed. Lai and Lam (1994) and Lai and Leung (1996) proposed various dispatching rules for yard cranes and tested them by simulation.

6 Strategic planning of multimodal systems

The focus of the models and methods presented in this section is broad: strategic planning issues at the international, national, and regional level, where the movements of several commodities through the transportation networks and services of several carriers are considered simultaneously. The main questions address the evolution of a given transportation system and its response to various modifications in its environment: evolution of the “local” or

international socioeconomic environment resulting in modifications to the patterns and volumes of production, consumption, and trade; modifications to existing policies and legislation and introduction of new ones (e.g., environment-related taxes); changes to existing infrastructure; variations in energy prices; modifications to labor conditions; carrier mergers; introduction of new technologies, and so on and so forth. These questions are often part of cost-benefit analyzes and comparative studies of policy and investment alternatives.

A strategic planning methodology identifies and represents the fundamental components of a transportation system – demand, supply, performance measures, and decision criteria – and their interactions. It models flow volumes by commodity and transportation mode, as well as associated performance measures, defined on a network representation of the transportation system. It aims to achieve a sufficiently good simulation of the global behavior of the system to offer a correct representation of the current situation and serve as an adequate analysis tool for planned or forecast scenarios and policies. It must be tractable and produce results that are easily accessible. No single formulation may address such a broad scope. Consequently, a strategic planning methodology is typically a set of models and procedures. Other than data manipulation tools (e.g., collection, fusion, updating, validation, etc.) and result analysis capabilities (e.g., cost-benefit, environmental impacts, energy consumption policies, etc.), the main components are:

(1) *Supply modeling* to represent the transportation modes, infrastructure, carriers, services, and lines; vehicles and convoys; terminals and intermodal facilities; capacities and congestion; economic, service, and performance measures and criteria.

(2) *Demand modeling* to capture the product definitions, identify producers, shippers, and intermediaries, and represent production, consumption, and zone-to-zone (region-to-region) distribution volumes, as well as *mode choices* for transportation. Relations of demand and mode choice to the performance of economic policies and transportation system performance are also addressed here.

(3) *Assignment* of multicommodity flows (from the demand model) to the multimode network (the supply representation). This procedure simulates the behavior of the transportation system and its output forms the basis for the strategic analyses and planning activities. The assignment methodology must therefore be both precise in reproducing current situations and sufficiently general to produce robust analyzes of future scenarios based on forecast data.

The prediction of multicommodity freight flows over a multimodal network is an important component of transportation science and has generated significant interest in recent years. However, perhaps due to the inherent difficulty and complexity of such problems, the study of freight flows at the national or regional level has not yet achieved full maturity, in contrast to passenger transportation where the prediction of car and transit flows over multimodal networks has been studied extensively and several of the research results have been transferred to practice (Florian and Hearn, 1995;

Cascetta, 2001). In the following, we review the most frequently used methodologies for freight planning and briefly describe associated references.

Demand

The modeling of demand attempts to describe the economic activities of a region, its production, consumption, import and export of goods. For planning purposes, the output of demand models is a series of product (or commodity group) specific demand matrices indicating the volumes to be moved from one zone to another. The process is often completed by the modeling of mode choice, which specifies for each product and origin–destination combination on what set of transportation infrastructure or services the demand may be moved.

A number of countries have developed *input/output* models of their economy that serve to determine the basic production and attraction of goods (Isard, 1951; Cascetta, 2001 and references within). In order to use an input/output model as a demand model, it is necessary to disaggregate the input/output model inputs and outputs by region and zone. This process is complex, and is usually not integrated with a supply representation and assignment procedure. When an input/output model is not available, the initial determination of origin–destination matrices is carried out by using national statistics on production, consumption, imports and exports combined with surveys of particular industrial sectors to complete missing or unreliable information. This process may be tedious since one has to reconcile data from several sources that may be collected by using different geographical subdivisions or inconsistent product definitions. The results of the disaggregated input/output model or the ad-hoc estimation procedures serve for the initial computation of origin-destination matrices for each product but without a subdivision by mode.

A second class of models that is well studied for the prediction of interregional commodity flows is the *spatial price equilibrium model* and its variants (Friesz et al., 1983; Harker and Friesz, 1986a, 1986b; Harker, 1987; see also Florian and Hearn, 1995; Nagurney, 1993). This class of models determines simultaneously the flows between producing and consuming regions, as well as the selling and buying prices that satisfy the spatial equilibrium conditions. A spatial equilibrium is reached when for all pairs of supply and demand regions with a positive commodity flow, the unit supply price plus the unit transportation cost is equal to the unit demand price; the sum is larger than this price for all pairs of regions with no exchanges. A simple network (bipartite graph) is generally used to represent the transportation system. These models rely to a large extent on the supply and demand functions of producers and consumers, respectively, which are rarely available and quite difficult to calibrate. There are relatively few applications of this class of models for the determination of demand by product and these deal with specific commodities such as crude oil, coal or milk products.

Mode choice

Mode-choice models aim to describe the set of transportation modes or services that may be used to carry specific products of groups thereof. The mode-choice definition may be rather general, e.g., petroleum moves by ship and pipeline, extremely specific indicating a particular set of single or multimodal paths for a given product, shipper, and origin–destination pair, or anywhere in between. The level of detail of modal specification need not be the same for all products or interzonal trade flows. The specification of mode choice may be inferred from historical data and shipper surveys or it may result from a formal description and modeling effort (Winston, 1983). The output of this process are either coefficients that indicate how to split the demand of a given origin–destination pair between the paths of a given set, or origin–destination demand matrices with particular sets of allowed modes.

Random utility models, developed and largely used for the analysis and planning of person transportation systems, have been proposed for freight transportation as well but their use in actual applications is scarce (Cascetta, 2001). The huge number of paths that have to be explicitly generated and stored, coupled to the challenge of performing this task for forecast data, may explain this phenomenon. At aggregated levels, mode choices have been specified for particularly important product flows by explicitly surveying the major logistic chains used between pairs of macro regions.

Supply representation and assignment

Once modal origin–destination matrices have been developed by some means, the next step is to assign them to the supply network model by using some route choice mechanism. The results of such an assignment model – product flows and performance measures – form part of the input to demand and cost-benefit modeling and analysis.

One class of assignment mechanisms is based again on the application of *random utility models* to the choice of paths defined previously by the mode-choice phase. It is noteworthy that the attributes of pre-defined paths are determined by the state of the network at generation time and are not responsive to assignment results. Thus, for example, congestion conditions are very difficult to represent. Moreover, the utility and choice models have to be calibrated, and all paths have to be generated, for each scenario, which is quite difficult to perform when forecast data is used.

Another class, *network optimization models* enable the prediction of multicommodity flows over a multimodal network that represents the transportation facilities at a level of detail appropriate for a nation or region. The demand and mode choice are exogenous and intermodal shipments are permitted. Within the specified mode choice, the optimization (assignment) engine determines the best (with respect to the specified network performance measures) multimodal paths for each product and origin–destination pair. The emphasis is

on the proper representation of the network and its different transportation modes, the corresponding intermodal transfer operations, the various criteria used to determine the movement of freight, the interactions and competition for limited resources captured through the representation of congestion effects, and the associated estimation of the traffic distribution over the transportation system considered to be used for comparative studies or for discrete time multiperiod analyses.

Studies in the 1970s used rather simple network representations (e.g., Jones and Sharp, 1977; Sharp, 1979). Several studies also attempted to extend spatial equilibrium models to include more refined network representations and to consider congestion effects and shipper–carrier interactions. Friesz et al. (1986) present a sequential model which uses two network representations: detailed separate networks for each carrier, and an aggregate, shipper-perceived network. On each carrier network commodities are transported at the least total cost. On the shipper-perceived network, traffic equilibrium principles are used to determine the carriers that shippers choose to move their traffic. This approach was quite successful in the study of logistics of products where a very limited number of shippers and carriers interact and strongly determine the behavior of the system. A typical example is the coal market between electric utilities in the United States and their suppliers in exporting countries. Friesz and Harker (1985), Harker and Friesz (1986a, 1986b), Harker (1987, 1988), and Hurley and Petersen (1994) present more elaborate formulations. This line of research has not, however, yielded practical planning models and tools yet, mainly because the formulations become too large and complex when applied to realistic situations. For a more detailed review of these efforts see Guélat et al. (1990) and Crainic et al. (1990b).

Models based on more sophisticated representations of the supply network were introduced by Guélat et al. (1990) and Jourquin and Beuthe (1996). We follow the former in the following presentation. The proposed modeling framework is that of a multimodal network, made up of modes, nodes, links, and intermodal transfers, on which multiple products are to be moved by specific vehicles and convoys between given origin and destination points. Here, a mode is a means of transportation having its own characteristics, such as vehicle type and capacity, as well as specific cost measures. Depending on the scope and level of detail of the strategic study, a mode may represent a carrier or part of its network representing a particular transportation service, an aggregation of several carrier networks, or specific transportation infrastructures such as ports.

The network consists of nodes $\mathcal{N}$, links $\mathcal{A}$, modes $\mathcal{M}$, and transfers $\mathcal{T}$ that represent all possible physical movements on the available infrastructure. To capture the modal characteristics of transportation, a link $a \in \mathcal{A}$ is defined as a triplet (i, m, j), where $i \in \mathcal{N}$ is the origin node, $j \in \mathcal{N}$ is the destination node, and $m \in \mathcal{M}$ is the mode. Parallel links are used to represent situations where more than one mode is available for transporting goods between two adjacent nodes. This network representation enables easy identification of the flow of

goods by mode, as well as various cost functions (e.g., operating cost, time delay, energy consumption, emissions, noise, risk, etc.) by product and mode. To model intermodal shipments, one must allow for mode transfers at certain nodes of the network and compute the associated costs and delays. Intermodal transfers $t \in \mathcal{T}$ at a node of the network are modeled as link to link, hence mode to mode, allowed movements. A path in this network then consists of a sequence of directed links of a mode, a possible transfer to another mode, a sequence of directed links of the second mode, and so on.

A product is any commodity (or collection of similar products) – goods or passengers – that generates a link flow. Each product $p \in \mathcal{P}$ transported over the multimodal network is shipped from certain origins $o \in \mathcal{N}$ to certain destinations $d \in \mathcal{N}$ within the network. The demand for each product is exogenous and is specified by a set of O–D matrices. The mode choice for each product is also exogenous and is indicated by defining for each O–D matrix a subset of modes allowed for transporting the corresponding demand. Shipper behavior is assumed to be reflected in these O–D matrices and associated mode choice. Let $g^{m(p)}$ be a demand matrix associated with product $p \in \mathcal{P}$, where $m(p) \subseteq \mathcal{M}$ is the subset of modes that may be used to move this particular part of product p.

The flows of product $p \in \mathcal{P}$ on the multimodal network are the decision variables of the model. Flows on links $a \in \mathcal{A}$ are denoted by v_a^p and flows on transfers $t \in \mathcal{T}$ are denoted by v_t^p; v stands for the vector of all product flows. Vehicle and convoy (e.g., train) movements are deduced from these flows. Cost functions are associated with the links and transfers of the network. For product p, the respective average cost functions $s_a^p(v)$ and $s_t^p(v)$ depend on the transported volume of goods. Then, the total cost of product p on arc a is $s_a^p(v)v_a^p$, and it is $s_t^p(v)v_t^p$ on transfer t. The total cost over the multimodal network is the function F, which is to be minimized over the set of flow volumes that satisfy the flow conservation and non negativity constraints:

$$F = \sum_{p \in P} \left(\sum_{a \in \mathcal{A}} s_a^p(v) v_a^p + \sum_{t \in \mathcal{T}} s_t^p(v) v_t^p \right). \tag{103}$$

Let $\mathcal{L}_{od}^{m(p)}$ denote the set of paths that for product p lead from origin o to destination d using only modes in $m(p)$. The path formulation of the flow conservation equations are then:

$$\sum_{l \in \mathcal{L}_{od}^{m(p)}} h_l = g_{od}^{m(p)}, \quad o, d \in \mathcal{N},\ p \in \mathcal{P},\ m(p) \subseteq \mathcal{M}, \tag{104}$$

where h_l is the flow on path $l \in \mathcal{L}_{od}^{m(p)}$. These constraints specify that the total flow moved over all the paths that may be used to transport product p must be equal to the demand for that product. The nonnegativity constraints are:

$$h_l \geqslant 0, \quad l \in \mathcal{L}_{od}^{m(p)},\ o, d \in \mathcal{N},\ p \in \mathcal{P},\ m(p) \subseteq \mathcal{M}. \tag{105}$$

The relation between arc flows and path flows is $v_a^p = \sum_{l \in \mathcal{L}^p} \delta_{al} h_l$, $a \in \mathcal{A}$, $p \in \mathcal{P}$, where $\mathcal{L}^p$ is the set of all paths that may be used by product p, and $\delta_{al} = 1$ if $a \in l$ (and 0, otherwise) is the indicator function which identifies the arcs of a particular path. Similarly, the flows on transfers are $v_t^p = \sum_{l \in \mathcal{L}^p} \delta_{tl} h_l$, $t \in \mathcal{T}$, $p \in \mathcal{P}$, where $\delta_{al} = 1$ if $t \in l$ (and 0, otherwise). Then, the system optimal multiproduct, multimodal assignment model consists of minimizing (103), subject to constraints (104) and (105). The optimality principle ensures that in the final flow distribution, for each product, demand matrix, and origin–destination pair, all paths with positive flows will have the same marginal cost (lower than on the other paths). The algorithm developed for this problem exploits the natural decomposition by product and results in a Gauss–Seidel-like procedure which allows the solution of large size problems in reasonable computational times (Guélat et al., 1990).

This model and algorithm are embedded in the STAN interactive-graphic system where they are complemented by a large number of tools to input, display, analyze, modify, and output data, as well as implement demand, mode choice, performance, and analysis models. See Larin et al. (2000) for a detailed description of the STAN system, components, interfaces, and tools. STAN has been applied successfully in practice for scenario analysis and planning, and several agencies and organizations in a number of countries around the world use it (Crainic et al., 1990a, 1990b, 1994, 1999, 2002; Guélat et al., 1990).

7 Perspectives

Intermodal transportation, particularly container-based, is steadily growing and will continue to do so in the foreseeable future. This is accompanied by the evolution of the regulatory, economic, and technological environment of the industry. Enhanced planning and management procedures and decision technologies are thus required, offering both great opportunities and significant challenges for the Operations Research community. On a general note, while many significant methodological advances have been achieved and several have been successfully transferred to actual practice, many problems have received scant attention. Moreover, advances in vehicle, infrastructure, and communication technologies yield new problems and require that problems already studied be revisited. In this section, we identify some of these trends, opportunities, and challenges.

Container terminals, mainly located in ports, are a case in point. Most research dedicated to this area is very recent and aimed at operational issues. This may be explained by the fact that terminals are often seen as bottlenecks in freight transportation and efforts are therefore dedicated to improving their efficiency and productivity. The development of comprehensive models for strategic and tactical planning of container terminals offers significant research opportunities. Moreover, the trend one observes in container terminal automation makes this research direction extremely timely. How to represent

automatic operations in planning models raises interesting questions, however. The automation of container terminals also opens up research opportunities in real-time decision and control of operations. Automated equipment collects and transmits data in realtime. This data, together with historical information and the plan of operations, could be used to automate and, in some case nearly optimize, real-time decisions. New models are required, as well as appropriate solution methods. Automation of terminals also requires revisiting operational planning models.

Compared to terminals, more work has been dedicated to carrier strategic and tactical planning issues. Yet, new problems emerge and many challenging research opportunities exist. Enhancing the models to better represent operation characteristics and to better integrate line and terminal activities is such an opportunity; integrating a representation of resource (vehicles, power, manpower) circulation and scheduling is another. The planning and operations of the "new" rail intermodal-service networks, operating regular and fixed services on a full-asset-utilization basis and enforcing advance bookings, define new challenging problems for Operations Research and Transportation Science.

The more comprehensive integration of the time-dependency of decisions and of the stochasticity of data and operations into strategic/tactic models is a major research challenge and opportunity. Indeed, many models aimed at strategic and tactical planning issues are static and almost all are deterministic. The conventional wisdom seems to be that such models plan based on "average" forecast data, while actual operations provide the "recourse" to adjust the plan to the day-to-day reality. Are we missing something? Could we build better plans and operate more efficiently by taking stochasticity into account at planning level? These and similar questions open up a fascinating research field.

Many operational problems are represented as time-dependent formulations. Crew scheduling, for example. Some, such as the management of empty vehicles, are also cast as stochastic formulations. Research opportunities exist in the development and enhancement of such formulations in many sectors of the industry. Thus, for example, very few results have been reported relative to the container fleet management problem. Moreover, each of these efforts was dedicated to one part of the system only. No comprehensive model of container fleet management on land and ocean is known to the authors. Moreover, while the scheduling and operation of various resources – vehicles, traction units, crews, etc. – are clearly related, most current formulations consider them separately. Research is required in this area.

The growth in the deployment of *Intelligent Transportation Systems (ITS)* and the electronic society will continue to impact the planning and operations of freight transportation. ITS and e-business technologies and procedures increase the flow of data, improve the timeliness and quality of information, and offer the possibility to control and coordinate operations in real-time. Research is required to adequately model the various planning and management

problems under ITS and real-time information and to develop efficient solution methods. These efforts must target carriers, terminals, as well as the entire intermodal chain. The scheduling, assignment, dispatching, routing, and re-routing of equipment are obvious and challenging subjects. As important is the impact on planning. Consider, to illustrate, the uncertainties related to the exact manifest of arriving container ships, their exact arrival time, and the destination (and, eventually, the carrier contracted for the next leg of the journey) of each container. ITS technology will deliver precise information earlier thus significantly reducing the uncertainty for the managers of the terminal and of the carriers that are next in the intermodal chain. Will uncertainty disappear completely? Unlikely. But its representation in planning and operations models will change.

Very few efforts have been dedicated to the intermodal chain. The e-business environment forces the issue and ITS offers the technological support. This largely unexplored field offers numerous research opportunities and challenges. The coordination (synchronization, in some cases) of plans and operations of independently owned or managed carriers and terminals is such a case. The uncertainties related to the operation of each element of the chain, the relations among these uncertainties, as well as their propagation within the intermodal chain are of prime importance in this context and pose considerable modeling and algorithmic challenges.

Recent years have brought to the forefront security issues related to transportation, ports, and border crossing. Planning and operations models and methods must be revisited and new ones must be proposed to address these issues, for each participant in the intermodal chain, as well as for the entire chain.

Most problems mentioned in this chapter are NP-hard and the formulations proposed are large-scale, mixed-integer combinatorial models. Stochastic, time-dependent formulations make resolution efforts even more difficult. And the need to build more comprehensive models is not making them any easier to solve. Significant research must, thus, be dedicated to the methodological aspects, including the study of models to develop stronger formulations and bounds. From an algorithmic point of view, the profession must continue to aim for powerful exact methods to address continuously larger problem instances. The dimension and complexity of the problems we face continuously overcome the capability of exact solution methods, however. Approximate solution methods, metaheuristics in particular, thus, play an increasingly important role in obtaining good solutions to difficult problems within reasonable computing times. Much work is still needed to develop more efficient and robust procedures and to better understand the conditions under which each method performs best. Solution methods that combine characteristics of two or more metaheuristics in a sequential or parallel computing setting offer interesting, but challenging perspectives.

Parallel and distributed computation indeed offers interesting perspectives with potentially great rewards: to solve realistically modeled and dimensioned

problem instances within reasonable times. Each class of problems and algorithms presents its own challenges. Promising research areas are the parallel exploration of branch-and-bound trees, the collaborative search undertaken by several metaheuristics or by metaheuristics and exact methods, and the development of hierarchical methods that combine different parallel models and methods (e.g., a first level of cooperating metaheuristics and branch-and-bound that each call, at a lower level, parallel procedures to evaluate solutions or bounds). Advanced decomposition techniques are also required, particularly related to the resolution of time-dependent problems. Parallel computing offers the possibility to design solution architectures to efficiently address complex requests in real, or quasi-real time. These ideas have just begun to be considered, but present great potential for the development of intelligent and efficient decision support tools for real-time intermodal transportation systems.

Acknowledgements

Funding for this project has been provided by the Natural Sciences and Engineering Council of Canada, through its Discovery Grant program. This work was supported in part by the Regional Research Centers Program (Research Center for Logistics Information Technology) granted by the Korean Ministry of Education & Human Resources Development.

While working on this project, Dr. T.G. Crainic was Adjunct Professor at the Département d'informatique et de recherche opérationnelle of the Université de Montréal (Canada) and at Molde University College (Norway).

References

Addinnour-Helm, S., Venkataramanan, M.A. (1998). Solution approaches to hub location problems. *Annals of Operations Research* 78, 31–50.

Ahuja, R.K., Magnanti, T.L., Orlin, J.B. (1993). *Network Flows – Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Aikens, C.H. (1985). Facility location models for distribution planning. *European Journal of Operational Research* 22, 263–279.

Alessandri, A., Sacone, S., Siri, S. (2006). Modelling and optimal receding-horizon control of maritime container terminals. *Journal of Mathematical Modelling and Algorithms*, in press.

Armacost, A.P., Barnhart, C., Ware, K.A. (2002). Composite variable formulations for express shipment service network design. *Transportation Science* 36 (1), 1–20.

Armacost, A.P., Barnhart, C., Ware, K.A., Wilson, A.M. (2004). UPS optimizes its air network. *Interfaces* 34 (1), 15–25.

Assad, A.A. (1980). Models for rail transportation. *Transportation Research Part A: Policy and Practice* 14, 205–220.

Aykin, T. (1990). On a quadratic integer program for the location of interacting hub facilities. *European Journal of Operational Research* 46 (3), 409–411.

Aykin, T. (1994). Lagrangian relaxation based approaches to capacitated hub-and-spoke network design problem. *European Journal of Operational Research* 79 (3), 501–523.

Aykin, T. (1995a). The hub location and routing problem. *European Journal of Operational Research* 83 (1), 200–219.

Aykin, T. (1995b). Networking policies for hub-and-spoke systems with application to the air transportation system. *Transportation Science* 29 (3), 201–221.

Balakrishnan, A., Magnanti, T.L., Mirchandani, P. (1997). Network design. In: Dell'Amico, M., Maffioli, F., Martello, S. (Eds.), *Annotated Bibliographies in Combinatorial Optimization*. Wiley, New York, pp. 311–334.

Ball, M., Magnanti, T.L., Monma, C.L., Nemhauser, G.L. (Eds.) (1995). *Network Routing*. *Handbooks in Operations Research and Management Science*, vol. 8. North-Holland, Amsterdam.

Barnhart, C., Schneur, R.R. (1996). Network design for express freight service. *Operations Research* 44 (6), 852–863.

Barnhart, C., Johnson, E.L., Nemhauser, G.L., Savelsbergh, M.W.F., Vance, P.H. (1998). Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46 (3), 316–329.

Barnhart, C., Hane, C.A., Vance, P.H. (2000). Using branch-and-price-and-cut to solve origin–destination integer multicommodity flow problems. *Operations Research* 48 (2), 318–326.

Bish, E.K. (2003). A multiple-crane-constrained scheduling problem in a container terminal. *European Journal of Operational Research* 144 (1), 83–107.

Bish, E.K., Chen, F.Y., Leong, Y.T., Nelson, B.L., Ng, J.W.C., Simchi-Levi, D. (2005). Dispatching vehicles in a mega container terminal. *OR Spectrum* 27 (4), 491–506.

Böse, J., Reiners, T., Steenken, D., Voß, S. (2000). Vehicle dispatching at seaport container terminals using evolutionary algorithms. In: *Proceedings of the 33rd Hawaii International Conference on System Sciences, vol. 2*. IEEE Computer Society, p. 2025.

Bostel, N., Dejax, P. (1998). Models and algorithms for container allocation problems on trains in a rapid transshipment shunting yard. *Transportation Science* 32 (4), 370–379.

Bourbeau, B., Gendron, B., Crainic, T.G. (2000). Branch-and-bound parallelization strategies applied to a depot location and container fleet management problem. *Parallel Computing* 26 (1), 27–46.

Braklow, J.W., Graham, W.W., Hassler, S.M., Peck, K.E., Powell, W.B. (1992). Interactive optimization improves service and performance for Yellow Freight system. *Interfaces* 22 (1), 147–172.

Brown, G.G., Lawphongpanich, S., Thurman, K.P. (1994). Optimizing ship berthing. *Naval Research Logistics* 41, 1–15.

Brown, G.G., Cormican, K.J., Lawphongpanich, S., Widdis, D.B. (1997). Optimizing submarine berthing with a persistence incentive. *Naval Research Logistics* 44, 301–318.

Buedenbender, K., Grünert, T., Sebastian, H.-J. (2000). A hybrid tabu search/branch and bound algorithm for the direct flight network design problem. *Transportation Science* 34 (4), 364–380.

Campbell, J.F. (1994a). Integer programming formulations of discrete hub location problems. *European Journal of Operational Research* 72, 387–405.

Campbell, J.F. (1994b). A survey of network hub location problems. *Studies in Locational Analysis* 6, 31–49.

Campbell, J.F. (1996). Hub location and the p-hub median problem. *Operations Research* 44 (6), 923–935.

Campbell, J.F., Ernst, A.T., Krishnamoorthy, M. (2002). Hub location problems. In: Drezner, Z., Hamacher, H. (Eds.), *Facility Location: Application and Theory*. Springer-Verlag, Berlin, pp. 373–407.

Cao, B., Uebe, G. (1995). Solving transportation problems with nonlinear side constraints with tabu search. *Computers & Operations Research* 22 (6), 593–603.

Cascetta, E. (2001). *Transportation Systems Engineering: Theory and Methods*. Kluwer Academic, Dordrecht, The Netherlands.

Castilho, B.D., Daganzo, C.F. (1991). Optimal pricing policies for temporary storage at ports. *Transportation Research Record* 1313, 66–74.

Castilho, B.D., Daganzo, C.F. (1993). Handling strategies for import containers at marine terminals. *Transportation Research Part B: Methodological* 27 (2), 151–166.

Chen, T. (1999). Yard operations in the container terminal – a study in the unproductive moves. *Maritime Policy Management* 26 (1), 27–38.

Cheung, R.K., Chen, C.-Y. (1998). A two-stage stochastic network model and solution methods for the dynamic empty container allocation problem. *Transportation Science* 32 (2), 142–162.
Cheung, R.K., Li, C.L., Lin, W. (2002). Interblock crane deployment in container terminals. *Transportation Science* 36 (1), 79–93.
Christiansen, M., Fagerholt, K., Ronen, D. (2004). Ship routing and scheduling: Status and perspectives. *Transportation Science* 38 (1), 1–18.
Christiansen, M., Fagerholt, K., Nygreen, B., Ronen, D. (2007). Maritime transportation. In: Barnhart, C., Laporte, G. (Eds.), *Transportation*. *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, pp. 189–284. This volume.
Cojeen, H.P., Dyke, P.V. (1976). The automatic planning and sequencing of containers for containership loading and unloading. In: Pitkin, M., Roche, J.J., Williams, T.J. (Eds.), *Ship Operation Automation, II: Proceedings of the 2nd IFAC/IFIP Symposium*. North-Holland, Amsterdam, pp. 415–423.
Cordeau, J.-F., Toth, P., Vigo, D. (1998). A survey of optimization models for train routing and scheduling. *Transportation Science* 32 (4), 380–404.
Cordeau, J.-F., Laporte, G., Legato, P., Moccia, L. (2005). Models and tabu search heuristics for the berth allocation problem. *Transportation Science* 39 (4), 526–538.
Cordeau, J.-F., Laporte, G., Savelsbergh, M.W.P., Vigo, D. (2007). Vehicle routing. In: Barnhart, C., Laporte, G. (Eds.), *Transportation*. *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, pp. 367–428. This volume.
Crainic, T.G. (1988). Rail tactical planning: Issues, models and tools. In: Bianco, L., Bella, A.L. (Eds.), *Freight Transport Planning and Logistics*. Springer-Verlag, Berlin, pp. 463–509.
Crainic, T.G. (2000). Network design in freight transportation. *European Journal of Operational Research* 122 (2), 272–288.
Crainic, T.G. (2003). Long-haul freight transportation. In: Hall, R.W. (Ed.), *Handbook of Transportation Science*, 2nd edition. Kluwer Academic, Norwell, MA, pp. 451–516.
Crainic, T.G., Delorme, L. (1995). Dual-ascent procedures for multicommodity location–allocation problems with balancing requirements. *Transportation Science* 27 (2), 90–101.
Crainic, T.G., Laporte, G. (1997). Planning models for freight transportation. *European Journal of Operational Research* 97 (3), 409–438.
Crainic, T.G., Rousseau, J.-M. (1986). Multicommodity, multimode freight transportation: A general modeling and algorithmic framework for the service network design problem. *Transportation Research Part B: Methodological* 20, 225–242.
Crainic, T.G., Ferland, J.-A., Rousseau, J.-M. (1984). A tactical planning model for rail freight transportation. *Transportation Science* 18 (2), 165–184.
Crainic, T.G., Dejax, P.J., Delorme, L. (1989). Models for multimode multicommodity location problems with interdepot balancing requirements. *Annals of Operations Research* 18, 279–302.
Crainic, T.G., Florian, M., Guélat, J., Spiess, H. (1990a). Strategic planning of freight transportation: STAN, an interactive-graphic system. *Transportation Research Record* 1283, 97–124.
Crainic, T.G., Florian, M., Léal, J.-E. (1990b). A model for the strategic planning of national freight transportation by rail. *Transportation Science* 24 (1), 1–24.
Crainic, T.G., Delorme, L., Dejax, P.J. (1993a). A branch-and-bound method for multicommodity location with balancing requirements. *European Journal of Operational Research* 65 (3), 368–382.
Crainic, T.G., Gendreau, M., Dejax, P.J. (1993b). Dynamic stochastic models for the allocation of empty containers. *Operations Research* 41 (1), 102–302.
Crainic, T.G., Gendreau, M., Soriano, P., Toulouse, M. (1993c). A tabu search procedure for multicommodity location/allocation with balancing requirements. *Annals of Operations Research* 41, 359–383.
Crainic, T.G., Florian, M., Larin, D. (1994). STAN: New developments. In: Khade, Al S., Brown, R. (Eds.), *Proceedings of the 23rd Annual Meeting of the Western Decision Sciences Institute*. School of Business Administration, California State University, Stanislaus, CA, pp. 493–498.
Crainic, T.G., Toulouse, M., Gendreau, M. (1995a). Parallel asynchronous tabu search for multicommodity location–allocation with balancing requirements. *Annals of Operations Research* 63, 277–299.
Crainic, T.G., Toulouse, M., Gendreau, M. (1995b). Synchronous tabu search parallelization strategies for multicommodity location-allocation with balancing requirements. *OR Spectrum* 17 (2–3), 113–123.

Crainic, T.G., Toulouse, M., Gendreau, M. (1997). Towards a taxonomy of parallel tabu search algorithms. *INFORMS Journal on Computing* 9 (1), 61–72.

Crainic, T.G., Dufour, G., Florian, M., Larin, D. (1999). Path analysis in STAN. In: Zopounidis, C., Despotis, D. (Eds.), *Proceedings 5th International Conference of the Decision Sciences Institute* New Technologies Publications, Athens, Greece, pp. 2060–2064.

Crainic, T.G., Dufour, G., Florian, M., Larin, D. (2002). Path recovery/reconstruction and applications in nonlinear multimodal multicommodity networks. In: Gendreau, M., Marcotte, P. (Eds.), *Transportation and Network Analysis: Current Trends: Miscellanea in Honor of Michael Florian*. Kluwer Academic, Norwell, MA, pp. 2060–2064.

Croxton, K.L., Gendron, B., Magnanti, T.L. (2003). Models and methods for merge-in-transit operations. *Transportation Science* 37 (1), 1–22.

Croxton, K.L., Gendron, B., Magnanti, T.L. (2006). Variable disaggregation in network flow problems with picewise linear costs. *Operation Research*, in press.

Daganzo, C.F. (1989a). Crane productivity and ship delay in ports. *Transportation Research Record* 1251, 1–9.

Daganzo, C.F. (1989b). The crane scheduling problem. *Transportation Research Part B: Methodological* 23 (3), 159–175.

Daganzo, C.F. (1990). The productivity of multipurpose seaport terminals. *Transportation Science* 24 (3), 205–216.

Daganzo, C.F. (2005). *Logistics Systems Analysis*, 4th edition. Springer-Verlag, Berlin.

Daskin, M.S. (1995). *Network and Discrete Location: Models, Algorithms, and Applications*. Wiley, New York.

Daskin, M.S., Owen, S.H. (2003). Location models in transportation. In: Hall, R.W. (Ed.), *Handbook of Transportation Science*, 2nd edition. Kluwer Academic, Norwell, MA, pp. 321–371.

Dejax, P.J., Crainic, T.G. (1987). A review of empty flows and fleet management models in freight transportation. *Transportation Science* 21 (4), 227–247.

Delorme, L., Roy, J., Rousseau, J.-M. (1988). Motor-carrier operation planning models: A state of the art. In: Bianco, L., Bella, A.L. (Eds.), *Freight Transport Planning and Logistics*. Springer-Verlag, Berlin, pp. 510–545.

Drezner, Z. (Ed.) (1995). *Facility Location: A Survey of Applications and Methods*. Springer-Verlag, New York.

Drezner, Z., Hamacher, H. (Eds.) (2002). *Facility Location: Application and Theory*. Springer-Verlag, Berlin.

Dror, M. (Ed.) (2000). *Arc Routing: Theory, Solutions and Applications*. Kluwer Academic, Norwell, MA.

Ebery, J., Krishnamoorthy, M., Ernst, A., Boland, N. (2000). The capacitated multiple allocation hub location problem: Formulations and algorithms. *European Journal of Operational Research* 120 (3), 614–631.

Equi, L., Gallo, G., Marziale, S., Weintraub, A. (1997). A combined transportation and scheduling problem. *European Journal of Operational Research* 97 (1), 94–104.

Erlenkotter, D. (1978). A dual-based procedure for uncapacitated facility location. *Operations Research* 26 (6), 992–1009.

Ermol'ev, Y.M., Krivets, T.A., Petukhov, V.S. (1976). Planning of shipping empty seaborne containers. *Cybernetics* 12, 644–646.

Ernst, A.T., Krishnamoorthy, M. (1996). Efficient algorithms for the uncapacitated single allocation p-hub median problem. *Location Science* 4 (3), 139–154.

European Conference of Ministers of Transport (2001). Terminology for combined transport. *http://www.crmt.org/online/glossaries/index.htm*.

Farvolden, J.M., Powell, W.B. (1991). A dynamic network model for less-than-truckload motor carrier operations. Working Paper 90-05, Department of Industrial Engineering, University of Toronto, Toronto, ON, Canada.

Farvolden, J.M., Powell, W.B. (1994). Subgradient methods for the service network design problem. *Transportation Science* 28 (3), 256–272.

Farvolden, J.M., Powell, W.B., Lustig, I.J. (1992). A primal partitioning solution for the arc–chain formulation of a multicommodity network flow problem. *Operations Research* 41 (4), 669–694.

Florian, M., Hearn, D. (1995). Network equilibrium models and algorithms. In: Ball, M., Magnanti, T.L., Monma, G.L., Nemhauser, C.L. (Eds.), *Network Routing*. *Handbooks in Operations Research and Management Science*, vol. 8. North-Holland, Amsterdam, pp. 485–550.

Fratar, T.J., Goodman, A.S., Brant Jr., A.E. (1960). Prediction of maximum practical berth occupancy. *Journal of the Waterways and Harbors Division*, 69–78.

Friesz, T.L., Harker, P.T. (1985). Freight network equilibrium: A review of the state of the art. In: Daughety, A.F. (Ed.), *Analytical Studies in Transport Economics*. Cambridge Univ. Press, Cambridge, Chapter 7.

Friesz, T.L., Tobin, R.L., Harker, P.T. (1983). Predictive intercity freight network models. *Transportation Research Part A: Policy and Practice* 17, 409–417.

Friesz, T.L., Gottfried, J.A., Morlok, E.K. (1986). A sequential shipper–carrier network model for predicting freight flows. *Transportation Science* 20, 80–91.

Gendron, B., Crainic, T.G. (1995). A branch-and-bound algorithm for depot location and container fleet management. *Location Science* 3 (1), 39–53.

Gendron, B., Crainic, T.G. (1997). A parallel branch-and-bound algorithm for multicommodity location with balancing requirements. *Computers & Operations Research* 24 (9), 829–847.

Gendron, B., Potvin, J.-Y., Soriano, P. (1999). Tabu search with exact neighbor evaluation for multicommodity location with balancing requirements. *INFOR* 37 (3), 255–270.

Gendron, B., Potvin, J.-Y., Soriano, P. (2003a). A parallel hybrid heuristic for the multicommodity capacitated location problem with balancing requirements. *Parallel Computing* 29, 591–606.

Gendron, B., Potvin, J.-Y., Soriano, P. (2003b). A tabu search with slope scaling for the multicommodity capacitated location problem with balancing requirements. *Annals of Operations Research* 122, 193–217.

Golden, B.L., Assad, A.A. (Eds.) (1988). *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam.

Griffiths, J.D. (1976). Optimal handling capacity at a berth. *Maritime Studies and Management* 3, 163–167.

Grünert, T., Sebastian, H.-J. (2000). Planning models for long-haul operations of postal and express shipment companies. *European Journal of Operational Research* 122 (2), 289–309.

Grünert, T., Sebastian, H.-J., Thäerigen, M. (1999). The design of a letter-mail transportation network by intelligent techniques. In: Sprague R. (Ed.), *Proceedings Hawaii International Conference on System Sciences 32*.

Grunow, M., Günther, H.-O., Lehmann, M. (2004). Dispatching multi-load AGVs in highly automated seaport container terminals. *OR Spectrum* 26 (2), 211–235.

Guan, Y., Cheung, R.K. (2004). The berth allocation problem: Models and solution methods. *OR Spectrum* 26 (1), 75–92.

Guan, Y., Xiao, W.Q., Cheung, R.K., Li, C.L. (2002). A multiprocessor task scheduling model for berth allocation: Heuristic and worst-case analysis. *Operations Research Letters* 30 (5), 343–350.

Guélat, J., Florian, M., Crainic, T.G. (1990). A multimode multiproduct network assignment model for strategic planning of freight flows. *Transportation Science* 24 (1), 25–39.

Günther, H.-O., Kim, K.H. (Eds.) (2005). *Container Terminals and Automated Transport Systems*. Springer-Verlag, Berlin.

Harker, P.T. (1987). *Predicting Intercity Freight Flows*. VNU Science Press, Utrech, The Netherlands.

Harker, P.T. (1988). Issues and models for planning and regulating freight transportation systems. In: Bianco, L., Bella, A.L. (Eds.), *Freight Transport Planning and Logistics*. Springer-Verlag, Berlin, pp. 374–408.

Harker, P.T., Friesz, T.L. (1986a). Prediction of intercity freight flows I: Theory. *Transportation Research Part B: Methodological* 20 (2), 139–153.

Harker, P.T., Friesz, T.L. (1986b). Prediction of intercity freight flows II: Mathematical formulations. *Transportation Research Part B: Methodological* 20 (2), 155–174.

Holguin-Veras, J., Jara-Diaz, S. (1999). Optimal pricing for priority service and space allocation in container ports. *Transportation Research Part B: Methodological* 33, 81–106.

Hurley, W.J., Petersen, E.R. (1994). Nonlinear tariffs and freight network equilibrium. *Transportation Science* 28 (3), 236–245.
Imai, A., Nagaiwa, K., Tat, C.W. (1997). Efficient planning of berth allocation for container terminals in Asia. *Journal of Advanced Transportation* 31 (1), 75–94.
Imai, A., Nishimura, E., Papadimitriou, S. (2001). The dynamic berth allocation problem for a container port. *Transportation Research Part B: Methodological* 35B (4), 401–417.
Imai, A., Nishimura, E., Papadimitriou, S. (2003). Berth allocation with service priority. *Transportation Research Part B: Methodological* 37 (5), 437–457.
Isard, W. (1951). Interregional and regional input–output analysis: A model of a space-economy. *The Review of Economics and Statistics* 33, 318–328.
ISL (2006). ISL shipping statistics and market review (SSMR) – short comment (6) 2006. Available at *http://www.isl.org*.
Jaillet, P., Song, G., Yu, G. (1996). Airline network design and hub location problems. *Location Science* 4 (3), 195–212.
Jones, P.S., Sharp, G.P. (1977). Multi-mode intercity freight transportation planning for underdeveloped regions. In: *Proceedings of the 18th Meeting of the Transportation Research Forum*, pp. 523–531.
Jordan, W.C., Turnquist, M.A. (1983). A stochastic dynamic network model for railroad car distribution. *Transportation Science* 17, 123–145.
Jourquin, B., Beuthe, M. (1996). Transportation policy analysis with a geographic information system: The virtual network of freight transportation in Europe. *Transportation Research Part C: Emerging Technologies* 4 (6), 359–371.
Kim, D., Barnhart, C., Ware, K., Reinhardt, G. (1999). Multimodal express package delivery: A service network design application. *Transportation Science* 33 (4), 391–407.
Kim, K.H. (1997). Evaluation of the number of rehandles in container yards. *Computers & Industrial Engineering* 32 (4), 701–711.
Kim, K.H., Bae, J.W. (1998). Remarshaling export containers in port container terminals. *Computers & Industrial Engineering* 35 (3–4), 655–658.
Kim, K.H., Bae, J.W. (1999). A dispatching method for automated guided vehicles to minimize delay of containership operations. *International Journal of Management Science* 5 (1), 1–25.
Kim, K.H., Bae, J.W. (2004). A look-ahead dispatching method for automated guided vehicles in automated port container terminals. *Transportation Science* 38 (2), 224–234.
Kim, K.H., Kim, H.B. (1999a). Segregating space allocation models for container inventories in port container terminals. *International Journal of Production Economics* 59, 415–423.
Kim, K.H., Kim, H.B. (2002). The optimal sizing of the storage space and handling facilities for import containers. *Transportation Research Part B: Methodological* 36, 821–835.
Kim, K.H., Kim, K.Y. (1999b). An optimal routing algorithm for a transfer crane in port container terminals. *Transportation Science* 33 (1), 17–33.
Kim, K.H., Kim, K.Y. (1999c). Routing straddle carriers for the loading operation of containers using a beam search algorithm. *Computers & Industrial Engineering* 36 (1), 109–136.
Kim, K.H., Moon, K.C. (2003). Berth scheduling by simulated annealing. *Transportation Research Part B: Methodological* 37 (6), 541–560.
Kim, K.H., Park, K.T. (2003a). A note on a dynamic space–allocation method for outbound containers. *European Journal of Operational Research* 148 (1), 92–101.
Kim, K.H., Park, K.T. (2003b). Dynamic space allocation for temporary storage. *International Journal of Systems Science* 34 (1), 11–20.
Kim, K.H., Park, Y.M. (2004). A crane scheduling method for port container terminals. *European Journal of Operational Research* 156 (3), 752–768.
Kim, K.H., Park, Y.M., Ryu, K.R. (2000). Deriving decision rules to locate export containers in container yard. *European Journal of Operational Research* 124 (1), 89–101.
Kim, K.H., Lee, K.M., Hwang, H. (2003). Sequencing delivery and receiving operations for yard cranes in port container terminals. *International Journal of Production Economics* 84 (3), 283–292.
Kim, K.H., Kang, J.S., Ryu, K.R. (2004). A beam search algorithm for the load sequencing of outbound containers in port container terminals. *OR Spectrum* 26 (1), 93–116.

Kim, K.Y., Kim, K.H. (1999d). A routing algorithm for a single straddle carrier to load export containers onto a containership. *International Journal of Production Economics* 59, 425–433.
Kim, K.Y., Kim, K.H. (2003). Heuristic algorithms for routing yard-side equipment for minimizing loading times in container terminals. *Naval Research Logistics* 50 (5), 498–514.
Klincewicz, J.G. (1990). Solving a freight transport problem using facility location techniques. *Operations Research* 38 (1), 99–109.
Klincewicz, J.G. (1991). Heuristics for the p-hub location problem. *European Journal of Operational Research* 53 (1), 25–37.
Klincewicz, J.G. (1992). Avoiding local optima in the p-hub location problem using tabu search and GRASP. *Annals of Operations Research* 40, 283–302.
Klincewicz, J.G. (1996). Dual algorithm for the uncapacitated hub location problem. *Location Science* 4 (3), 173–184.
Koh, Y.K., Kim, S.C. (2001). A study on the impact of extra large-sized containership on the shipping company & port. *Korean International Commerce* 16 (2), 165–187.
Kozan, E. (1997). Comparison of analytical and simulation planning models of sea port container terminals. *Transportation Planning and Technology* 20, 235–248.
Kozan, E. (2000). Optimizing container transfers at multimodal terminals. *Mathematical and Computer Modeling* 31, 235–243.
Kozan, E., Preston, P. (1999). Genetic algorithms to schedule container transfers at multimodal terminals. *International Transactions in Operational Research* 6, 311–329.
Kuby, M.J., Gray, R.G. (1993). The hub network design problem with stopovers and feeders: The case of Federal Express. *Transportation Research Part A: Policy and Practice* 27 (1), 1–12.
Labbé, M., Louveaux, F.V. (1997). Location problems. In: Dell'Amico, M., Maffioli, F., Martello, S. (Eds.), *Annotated Bibliographies in Combinatorial Optimization*. Wiley, New York, pp. 261–281.
Labbé, M., Peeters, D., Thisse, J.-F. (1995). Location on networks. In: Ball, M., Magnanti, T.L., Monma, G.L., Nemhauser, C.L. (Eds.), *Network Routing. Handbooks in Operations Research and Management Science*, vol. 8. North-Holland, Amsterdam, pp. 551–624.
Lai, K.K., Lam, K. (1994). A study of container yard equipment allocation strategy in Hong Kong. *International Journal of Modeling & Simulation* 14 (3), 134–138.
Lai, K.K., Leung, J.W. (1996). Analysis of yard crane deployment strategies in a container terminal. In: *Proceedings of ICC & IE Conference*, Kyungju, Korea, pp. 1187–1190.
Lai, K.K., Leung, J. (2000). Analysis of gate house operations in a container terminal. *International Journal of Modelling and Simulation* 20 (1), 89–94.
Lai, K.K., Shih, K. (1992). A study of container berth allocation. *Journal of Advanced Transportation* 26 (1), 45–60.
Larin, D., Crainic, T.G., Simonka, G., James-Lefebvre, L., Dufour, G., Florian, M. (2000). STAN user's manual, Release 6. INRO Consultants, Inc., Montréal, QC, Canada.
Li, C.-L., Cai, X., Lee, C.-Y. (1998). Scheduling with multiple-job-on-one-processor pattern. *IEE Transactions on Scheduling and Logistics* 30 (5), 433–445.
Lim, A. (1998). The berth planning problem. *Operations Research Letters* 22 (2–3), 105–110.
Macharis, C., Bontekoning, Y.M. (2004). Opportunities for OR in intermodal freight transport research: A review. *European Journal of Operational Research* 153 (2), 400–416.
Magnanti, T.L., Wong, R.T. (1984). Network design and transportation planning: Models and algorithms. *Transportation Science* 18 (1), 1–55.
Mattfeld, D.C., Kopfer, H. (2004). Terminal operations management in vehicle transshipment. *Transportation Research Part A: Policy and Practice* 37 (5), 435–452.
Miller, A.J. (1971). Queuing at single-berth shipping terminal. *Journal of Waterways, Harbors and Coastal Engineering Division* 97, 43–56.
Minoux, M. (1989). Network synthesis and optimum network design problems: Models, solution methods and applications. *Networks* 19, 313–360.
Mirchandani, P.S., Francis, R.L. (Eds.) (1990). *Discrete Location Theory*. Wiley, New York.
Nagurney, A. (1993). *Network Economics: A Variational Inequality Approach*. Kluwer Academic, Norwell, MA.

Nam, K.-C., Kwak, K.-S., Yu, M.-S. (2002). Simulation study of container terminal performance. *Journal of Waterway, Port, Coastal and Ocean Engineering* 128 (3), 126–132.

Narasimhan, A., Palekar, U.S. (2002). Analysis and algorithms for the transtainer routing problem in container port operations. *Transportation Science* 36 (1), 63–78.

Nemhauser, G.L., Wolsey, L.A. (1993). *Integer and Combinatorial Optimization*. Wiley, New York.

Noritake, M., Kimura, S. (1990). Optimum allocation and size of seaports. *Journal of Waterway, Port, Coastal, and Ocean Engineering* 116 (2), 287–299.

O'Kelly, M.E. (1987). A quadratic integer program for the location of interacting hub facilities. *European Journal of Operational Research* 32 (3), 393–404.

O'Kelly, M.E. (1992a). A clustering approach to the planar hub location problem. *Annals of Operations Research* 40, 339–353.

O'Kelly, M.E. (1992b). Hub facilities location with fixed costs. *Papers in Regional Science* 71, 292–306.

O'Kelly, M.E., Skorin-Kapov, D., Skorin-Kapov, J. (1995). Lower bounds for the hub location problem. *Management Science* 41 (4), 713–721.

O'Kelly, M.E., Bryan, D., Skorin-Kapov, D., Skorin-Kapov, J. (1996). Hub network design with single and multiple allocation: A computational study. *Location Science* 4 (3), 125–138.

Park, K.T., Kim, K.H. (2002). Berth scheduling for container terminals by using a subgradient optimization technique. *Journal of the Operational Research Society* 53 (9), 1049–1054.

Park, Y.M. (2003). Berth and crane scheduling of container terminals. Department of Industrial Engineering, Pusan National University, Pusan, Korea.

Park, Y.M., Kim, K.H. (2003). A scheduling method for berth and quay cranes. *OR Spectrum* 25 (1), 1–23.

Pedersen, M.B., Crainic, T.G., Madsen, O.B.G. (2006). Models and tabu search metaheuristics for service network design with vehicle balance requirements. Publication CRT-2006-22, Centre de recherche sur les transports, Université de Montréal, Montréal, QC, Canada.

Peterkofsky, R.I., Daganzo, C.F. (1990). A branch and bound solution method for the crane scheduling problem. *Transportation Research Part B: Methodological* 24 (3), 159–172.

Pirkul, H., Jayaraman, V. (1996). Production, transportation, and distribution planning in a multi-commodity tri-echelon system. *Transportation Science* 30 (4), 291–302.

Pirkul, H., Jayaraman, V. (1998). A multi-commodity, multi-plant, capacitated facility location problem: Formulation and efficient heuristic solution. *Computers & Operations Research* 25 (10), 869–878.

Powell, W.B. (1986). A local improvement heuristic for the design of less-than-truckload motor carrier networks. *Transportation Science* 20 (4), 246–357.

Powell, W.B. (1988). A comparative review of alternative algorithms for the dynamic vehicle allocation problem. In: Golden, B.L., Assad, A.A. (Eds.), *Vehicle Routing: Methods and Studies*. North-Holland, Amsterdam, pp. 249–292.

Powell, W.B. (2003). Dynamic models of transportation operations. In: Graves, S., Tok, T.A.G. (Eds.), *Supply Chain Management*. *Handbooks in Operations Research and Management Science*, vol. 11. North-Holland, Amsterdam, pp. 677–756.

Powell, W.B., Carvalho, T.A. (1998). Real-time optimization of containers and flatcars for intermodal operations. *Transportation Science* 32 (2), 110–126.

Powell, W.B., Sheffi, Y. (1983). The load-planning problem of motor carriers: Problem description and a proposed solution approach. *Transportation Research Part A: Policy and Practice* 17 (6), 471–480.

Powell, W.B., Sheffi, Y. (1986). Interactive optimization for motor carrier load planning. *Journal of Business Logistics* 7 (2), 64–90.

Powell, W.B., Sheffi, Y. (1989). Design and implementation of an interactive optimization system for the network design in the motor carrier industry. *Operations Research* 37 (1), 12–29.

Powell, W.B., Topaloglu, H. (2003). Stochastic programming in transportation and logistics. In: Ruszczynski, A., Shapiro, A. (Eds.), *Stochastic Programming*. *Handbooks in Operations Research and Management Science*, vol. 10. North-Holland, Amsterdam, pp. 555–635.

Powell, W.B., Topaloglu, H. (2005). Fleet management. In: Wallace, S., Ziemba, W. (Eds.), *Applications of Stochastic Programming*. *Math Programming Society–SIAM Series on Optimization*. SIAM, Philadelphia, PA, pp. 185–216.

Powell, W.B., Jaillet, P., Odoni, A. (1995). Stochastic and dynamic networks and routing. In: Ball, M., Magnanti, T.L., Monma, G.L., Nemhauser, C.L. (Eds.), *Network Routing*. *Handbooks in Operations Research and Management Science*, vol. 8. North-Holland, Amsterdam, pp. 141–295.

Powell, W.B., Bouzaïene-Ayari, B., Simaõ, H.P. (2007). Dynamic models for freight transportation. In: Barnhart, C., Laporte, G. (Eds.), *Transportation*. *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, pp. 285–365. This volume.

Preston, P., Kozan, E. (2001). An approach to determine storage locations of containers at seaport terminals. *Computers & Industrial Engineering* 28, 983–995.

Ramani, K.V. (1996). An interactive simulation model for the logistics planning of container operations in seaports. *Simulation* 66 (5), 291–300.

Roy, J., Crainic, T.G. (1992). Improving intercity freight routing with a tactical planning model. *Interfaces* 22 (3), 31–44.

Roy, J., Delorme, L. (1989). NETPLAN: A network optimization model for tactical planning in the less-than-truckload motor-carrier industry. *INFOR* 27 (1), 22–35.

Ryu, K.R., Kim, K.H., Lee, Y.H., Park, Y.M. (2001). Load sequencing algorithms for container ships by using metaheuristics. In: *Proceedings of 16th International Conference on Production Research*. Prague, Czech Republic, 29 July–4 August, CD-ROM.

Salkin, H.M., Mathur, K. (1989). *Foundations of Integer Programming*. North-Holland, Amsterdam.

Schonfeld, P., Sharafeldien, O. (1985). Optimal berth and crane combinations. *Journal of Waterway, Port, Coastal and Ocean Engineering* 111 (6), 1060–1072.

Sharp, G.P. (1979). A multi-commodity, intermodal transportation model. In: *Proceedings of the 20th Meeting of the Transportation Research Forum*, pp. 399–407.

Skorin-Kapov, D., Skorin-Kapov, J. (1994). On tabu search for the location of interacting hub facilities. *European Journal of Operational Research* 73 (3), 502–509.

Skorin-Kapov, D., Skorin-Kapov, J., O'Kelly, M.E. (1996). Tight linear programming relaxation of uncapacitated p-hub median problems. *European Journal of Operational Research* 94 (3), 582–593.

Smith, K., Krishnamoorthy, M., Palaniswami, M. (1996). Neural versus traditional approaches to the location of interacting hub facilities. *Location Science* 4 (3), 155–171.

Steenken, D., Voß, S., Stahlbock, R. (2004). Container terminal operation and operations research – a classification and literature review. *OR Spectrum* 26 (1), 3–49.

Taleb-Ibrahimi, M., Castilho, B., Daganzo, C.F. (1993). Storage space vs handling work in container terminals. *Transportation Research Part B: Methodological* 27 (1), 13–32.

Toth, P., Vigo, D. (Eds.) (2002). *The Vehicle Routing Problem*. *SIAM Monographs on Discrete Mathematics and Applications*. SIAM, Philadelphia, PA.

van der Heyden, W.P.A., Ottjes, J.A. (1985). A decision support system for the planning of the workload of a grain terminal. *Decision Support Systems* 1, 293–297.

van Hee, K.M., Wijbrands, R.J. (1988). Decision support system for container terminal planning. *European Journal of Operational Research* 34 (3), 262–272.

Vis, I.F.A., de Koster, R., Roodbergen, K.J. (2001). Determination of the number of automated guided vehicles required at a semi-automated container terminal. *Journal of the Operational Research Society* 52 (4), 409–417.

Wanhill, S.R.C. (1974). Further analysis of optimum size seaport. *Journal of the Waterways Harbors and Coastal Engineering Division* 100, 377–383.

White, W. (1972). Dynamic transshipment networks: An algorithm and its application to the distribution of empty containers. *Networks* 2 (3), 211–236.

Winston, C. (1983). The demand for freight transportation: Models and applications. *Transportation Research Part A: Policy and Practice* 17, 419–427.

Zhang, C., Wan, Y.W., Liu, J., Linn, R.J. (2002). Dynamic crane deployment in container storage yards. *Transportation Research Part B: Methodological* 36 (6), 537–555.

Zhang, C., Liu, J., Wan, Y.-W., Murty, K.G., Linn, R. (2003). Storage space allocation in container terminals. *Transportation Research Part B: Methodological* 37 (10), 883–903.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14009-8

Chapter 9

Hazardous Materials Transportation

Erhan Erkut

Faculty of Business Administration, Bilkent University, Ankara, Turkey
E-mail: erkut@bilkent.edu.tr

Stevanus A. Tjandra

University of Alberta School of Business, Edmonton, Canada
E-mail: Stevanus.Tjandra@ualberta.ca

Vedat Verter

Desautels Faculty of Management, McGill University, Montreal, Canada
E-mail: vedat.verter@mcgill.ca

1 Introduction

The transportation of hazardous materials (or dangerous goods) deserves to be treated in a separate chapter of this volume, primarily due to the risks associated with this activity. Although the industry has an excellent safety record, accidents do happen, and the consequences can be significant, due to the nature of the cargo. Reduction of hazardous material (hazmat) transportation risks can be achieved in many different ways. Some of these risk reduction measures, such as driver training and regular vehicle maintenance, have little connection to operations research (OR), whereas others offer interesting challenges to OR. This chapter focuses on applications of OR models to hazmat transportation, providing a relatively comprehensive review of the literature, and outlining areas of potential impact for operations researchers.

According to the US Department of Transportation (US DOT), a hazardous material is defined as any substance or material capable of causing harm to people, property, and the environment. Dependence on hazardous materials is a fact of life in industrialized societies. There are thousands of different hazardous materials in use today (US DOT, 2004b). The United Nations sorts hazardous materials into nine classes according to their physical, chemical, and nuclear properties: explosives and pyrotechnics; gasses; flammable and combustible liquids; flammable, combustible, and dangerous-when-wet solids; oxidizers and organic peroxides; poisonous and infectious materials; radioactive materials; corrosive materials (acidic or basic); and miscellaneous dangerous goods, such as hazardous wastes (UN, 2001). In almost all instances, hazmats originate at a location other than their destination. For example, oil is extracted from oil fields and shipped to a refinery (typically via a pipeline); many oil products, such as heating oil and gasoline, are refined at the refinery and then

shipped to storage tanks at different locations within a country. As another example, polychlorinated biphenyls (PCBs) are collected at many industrial installations, such as old power generation and transfer stations and shipped to a special waste management facility for safe disposal (usually incineration). Hence, transportation plays a significant role for hazmats. The magnitude of this role depends on the size of a country and its level of industrialization. For example, the Office of Hazardous Materials Safety (OHMS) of the US DOT estimated that there were 800,000 domestic shipments of hazmats, totaling approximately 9 million tons, in the USA each day in 1998 (US DOT, 2000). Transport Canada estimates that nearly 80,000 shipments of dangerous goods are moved by road, rail, water, and air in Canada (Transport Canada, 2004). Given a conservative estimate of 2% annual growth in the production of hazmats, it is safe to assume that the total number of shipments in North America is well over the one million mark in 2005.

In 2002, over 99 percent of hazmat shipments in Canada made it safely to their destination (Transport Canada, 2004). While the hazmat transport sector is far safer than other transport sectors (US DOT, 2000), hazmat transport accidents do happen. Figure 1 shows the distribution of accidents/incidents by hazmat class in 2003. An accident resulting in a release of the hazmat is called an incident. The figure shows that flammable–combustible liquids and corrosive materials accounted for the majority of hazmat accidents/incidents in the USA (US DOT, 2004a).

The transportation of hazmats can be classified according to the mode of transport, namely: road, rail, water, air, and pipeline. Some shipments are intermodal; they are switched from one mode to another during transit. There

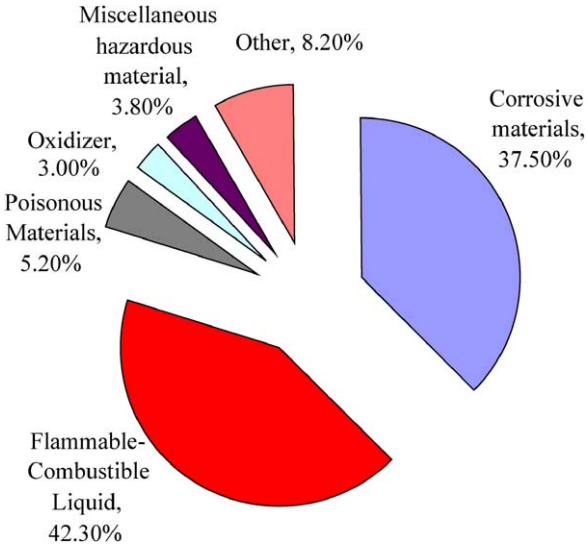


Fig. 1. Accident/incident by hazmat class in 2003 (US DOT, 2004a).

are significant differences in the use of these modes. While transportation by truck accounts for approximately 94% of all individual hazmat shipments in the USA, this mode carries merely 43% of the hazmat tonnage since the volume that can be shipped by one truck is limited compared to other modes of transport. In contrast, rail, water, and pipelines carry 57% of the hazmat tonnage while accounting for less than 1% of all individual shipments. It is possible to carry huge quantities of hazmats using these modes. While the counting of individual shipments is less clear with these modes (How do we count the number of shipments via a pipeline? Does a train consisting of multiple hazmat tank cars count as a single shipment?), they carry much larger quantities per shipment than trucks do. The balance of hazmat shipments (5% by count and 0.05% by weight) are made via air (US DOT, 1998).

Hazmat transport incidents can occur at the origin or destination (when loading and unloading) or en-route. Incidents involving hazmat cargo can lead to severe consequences characterized by fatalities, injuries, evacuation, property damage, environmental degradation, and traffic disruption. In 2003, there were 488 serious incidents (among a total of 15,178 incidents) resulting in 15 deaths, 17 major and 18 minor injuries, and a total property damage of $37.75 million (US DOT, 2004c). About 90% of hazmat incidents occur on highways. As far as causes go, human error seems to be the single greatest factor (see Figure 2) in all hazardous materials incidents (minor and serious incidents).

The annual number of nonhazmat transportation accidents in the USA is estimated to be 126,880, in contrast to the approximately 15,000 hazmat trans-

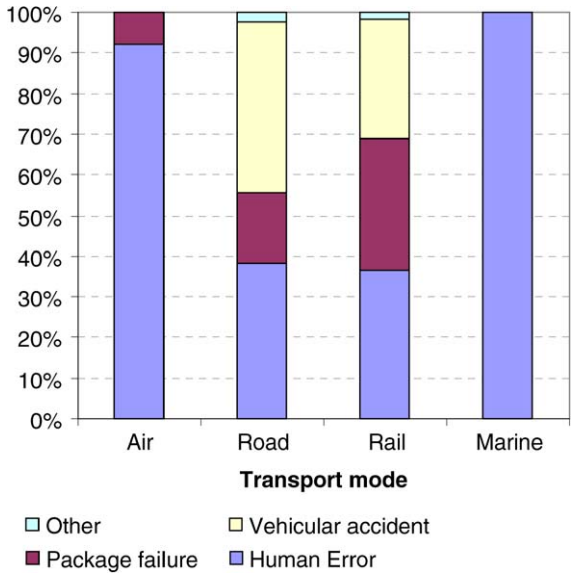


Fig. 2. Serious incident by mode/cause in 2003 (US DOT, 2004c).

portation accidents and incidents (FMCSA, 2001). Even though hazmats are involved in a small minority of all transport accidents, hazmat accidents can have catastrophic consequences. In 2003, for example, 22 train cars derailed at Tamoroa, IL, resulting in the release of various types and quantities of hazardous materials from seven tank cars. The evacuation of over a thousand residents within a three-mile radius and the closing of Highway 51 followed the derailment.

Table 1 contrasts the average costs (per event) of hazmat and nonhazmat motor carrier accidents and incidents for one year. Although the cost of an average hazmat incident is not significantly higher than the cost of a nonhazmat incident, the cost of a hazmat incident resulting in fire or explosion is significantly higher. Hazmat transportation accidents are perceived as low probability–high consequence (LPHC) events and data seem to support this perception. For example, chlorine leaking from damaged tank cars due to a derailment in Mississauga, Ontario in 1979, forced the evacuation of 200,000 people. In 1982, a gasoline truck explosion in a tunnel in Afghanistan caused 2700 fatalities. Most transport accidents that impact a large number of people and result in significant economic loss involve a hazmat cargo.

Hazmat transportation involves multiple players such as shippers, carriers, packaging manufacturers, freight forwarders, consignees, insurers, governments, and emergency responders; each has a different role in safely moving hazardous materials from their origins to their destinations. There are often multiple handoffs of material from one party to another during transport. The various parties, ranging from individuals or small firms to large multinational organizations, may have overlapping and unclear responsibilities for managing the risks (ICF Consulting, 2000). Furthermore, each party may have different priorities and viewpoints. Although the transportation department or local government is responsible for designating allowable routes that reduce risk, a carrier company would, in general, try to identify the route that minimizes the fuel costs and travel times, between the origin and destination for each shipment. Some routes have short lengths but move through heavily populated areas; some routes avoid heavily populated areas but are longer, resulting in higher transport costs and accident probabilities; while other routes use major

Table 1.
Comparative costs of hazmat and nonhazmat motor carrier accident/incident events (FMCSA, 2001)

Type of accident/incident event	Average cost (in US$)	Average traffic delay (in hours)
Nonhazmat events	340,000	2
All hazmat events	414,000	–
Hazmat events with spill/release	536,000	5
Hazmat events with fire	1,200,000	8
Hazmat events with explosion	2,100,000	12

freeways and thus minimize travel time but may be associated with higher accident rates. Thus, hazmat transportation is a typical multiobjective problem with multiple stakeholders.

Multiobjective/multistakeholder problems are complicated to solve. Hazmat transport problems are further complicated by public sensitivity surrounding these problems. The concept of *social amplification of risk* (see Kasperson et al., 1988; Renn et al., 1992) indicates that public assessment of a risk depends not only on its magnitude but also on subjective perceptions. The individual and social perceptions of risk can be heightened or attenuated by many factors such as extensive media coverage of the hazard event (see, e.g., Horlick-Jones, 1995), involvement of social groups (see, e.g., Moore, 1989), inaccuracies and inconsistencies in the communication process that lead to rumors and speculations on risk magnitude (see, e.g., Mileti and O'Brien, 1992; Barnes, 2001). The amplification of the risk of a relatively minor hazmat accident may imply much stronger public reaction and results in a call for action, such as tighter transport regulations or even the banning of hazmat shipments via a certain mode of transport, in some extreme cases.

Public sensitivity to hazmat transport is rooted not only in public risk perceptions, but also in equity concerns. Those individuals benefiting from hazmat shipments are usually those who live near the production facility or the delivery points. Yet the population living along a major highway connecting the hazmat origin and destinations is exposed to the transport risks regardless of whether or not they benefit from the hazmat shipments. This lack of burden-benefit concordance is another source of public opposition to hazmat shipments. The shipment of spent nuclear fuel rods from nuclear power plants to the proposed repository at Yucca Mountain in Nevada, USA, offers a good example of equity-based public opposition. The shipping reduces the risk at the power plants. Yet some risk is imposed on the population living along the major east–west highways or railways, who are asked to assume the risk with no clear benefits to them. Furthermore, if the same main route segment were selected for shipments from multiple origins, the objection of people living along this route would increase considerably. These people are likely to prefer alternate routings that would spread the risks.

Public opposition to hazmat shipments has increased in recent years, due to fears of terrorist attacks on hazmat vehicles. The Research and Special Projects Administration (RSPA) of US DOT accepts that hazmats could pose a significant threat during transportation, when they are particularly vulnerable to sabotage or misuse as weapons of mass destruction or as weapons of convenience by terrorists – particularly given how easy it is to identify a hazmat vehicle (as well as the specifics of their cargo) given the current system of hazmat placards. As a result some jurisdictions are trying to force a rerouting of hazmat vehicles away from populated areas by implementing local laws.

Much of the discussion to this point also applies to the location of hazardous facilities. If anything, the risks and the public opposition are higher for fixed facilities than for transport. Operations researchers have dealt with both types

of problems, and we will include references to facility location as well as transportation problems in this chapter, particularly for facility location models that treat the transportation component explicitly.

The rest of this chapter is organized in the following way. In Section 2, we offer a high-level view of hazmat logistics literature where we summarize special journal issues, reference books, reports, and web sites that are potentially useful to an operations researcher who wishes to conduct research in this area. We also offer a classification of journal papers, which provides the organizational structure for the rest of the chapter. Section 3 contains a treatment of risk, the main ingredient of hazmat logistics problems that separate them from other logistics problems. We review different models of risk for hazmat transport and discuss how one can go from point risk to edge risk and then to route risk. Section 4 deals with hazmat routing and scheduling problems. In Section 5, we turn our attention to models that combine undesirable facility location and hazmat transportation. In the final section we offer a critique of the existing literature and suggest directions for future research.

2 A high-level view of hazmat logistics research

2.1 Special issues of journals

Hazmat logistics has been a very active research area during the last twenty years. In 1984 *Management Science* published a special issue on *Risk Analysis* (Vol. 30, No. 4) where five papers dealt with hazmats and hazardous facilities. This was followed by a number of special issues of refereed academic journals that focus on hazmat transportation or location problems.

- *Transportation Research Record* published two special issues on hazmat transportation in 1988 (No. 1193) that included four papers and 1989 (No. 1245) that included six papers.
- *Transportation Science* devoted an issue to hazmat logistics in 1991 (Vol. 25, No. 2) that contained six papers.
- There was a special section on hazmat transportation in the March/April 1993 issue of the *Journal of Transportation Engineering* that included four papers.
- A special double-issue of *INFOR* on hazardous materials logistics was published in 1995 (Vol. 33, No. 1 and 2) with nine papers.
- Four papers were included in a special issue of *Location Science* dealing with hazmats in 1995 (Vol. 3, No. 3).
- *Transportation Science* produced a second special issue on hazmat logistics in 1997 (Vol. 31, No. 3) with seven papers.
- *Studies in Locational Analysis* published a special issue on undesirable facility location in April 1999 (Issue 12) that contained seven papers.

- *Computers & Operations Research* will publish a hazmat logistics special issue in 2007 which will contain results of the most recent research in the area in 13 papers.

These special issues contain many useful papers and they offer a good starting point for research in this area. Likewise, the book chapter by Erkut and Verter (1995a) offers a relatively comprehensive survey of the literature up to 1994, and the annotated bibliography by Verter and Erkut (1995) offers a good list of pre-1994 references in risk assessment, location, and routing.

2.2 Books

Perhaps an even better starting point for those who wish to familiarize themselves with the terminology and the problem context are the following books.

- *Transportation of Hazardous Materials: Issues in Law, Social Science, and Engineering* (1993), edited by L.N. Moses and D. Lindstrom, Kluwer Academic Publishers. This book contains 18 articles presented at *Hazmat Transport '91*, a national conference held at Northwestern University on all aspects of hazmat transport. While only a few of the articles use OR models and techniques, the book offers a multidimensional treatment of the subject and it is good reading for new researchers in the area.
- Three books were produced by Institute for Risk Research, University of Waterloo, as a result of the First International Consensus Conference on the Risks of Transporting Dangerous Goods, held in Toronto, Canada in April, 1992:
  - *Transportation of Dangerous Goods: Assessing the Risks* (1993), edited by F.F. Saccomanno and K. Cassidy. This book contains 30 articles which are organized into five main chapters: Application of QRA models to the transport of Dangerous Goods; Analysis of Dangerous Goods Accident and Releases; Application of Simple Risk Assessment Methodology; Uncertainty in Risk Estimation; Risk Tolerance, Communication and Policy Implications.
  - *Comparative Assessment of Risk Model Estimates for the Transport of Dangerous Goods by Road and Rail* (1993), edited by F.F. Saccomanno, D. Leming, and A. Stewart. This book documents the assessment of a corridor exercise involving the application of several risk models to a common transport problem involving the bulk shipment of chlorine, LPG, and gasoline by road and rail along predefined routes. The purpose of the corridor exercise was to provide a well-defined transportation problem for analysis in order to examine the sources of variability in the risk estimates. Seven agencies in six countries participated in this exercise.
  - *What is the Risk* (1993), edited by F.F. Saccomanno, D. Leming, and A. Stewart. This book documents the small group discussions and

consensus testing process from the corridor exercise conducted as part of the international consensus conference.

- *Hazardous Materials Transportation Risk Analysis: Quantitative Approaches for Truck and Train* (1994), Rhyne WR, Van Norstrand Reinhold. This book explains the quantitative risk analysis (QRA) methodologies and their application to hazmat transportation. It also provides an extended example of a QRA for bulk transport of chlorine by truck and train. This detailed example explores every step of the QRA from preliminary hazards analysis to risk reduction alternatives. This book is a valuable reference for hazmat transportation risks, and it is intended for practitioners. It is not an OR book, but it provides useful information for OR research in hazmat transportation modeling and analysis.
- *Guidelines for Chemical Transportation Risk Analysis* (1995), American Institute of Chemical Engineers, Center for Chemical Process Safety (CCPS), New York. This book completes two other books in the series of process safety guidelines books produced by CCPS: Guidelines for Chemical Process Quantitative Risk Analysis (CPQRA, 1989) and Guidelines for Hazard Evaluation Procedures (HEP, 2nd edition, 1992). It is intended to be used as a companion volume to the CPQRA and HEP Guidelines when dealing with a quantitative transportation risk analysis (TRA) methodology. This book offers a basic approach to TRA for different transport modes (pipelines, rail, road, barge, water, and intermodal containers). It can be useful to an engineer or manager in identifying cost effective ways to manage and reduce the risk of a hazmat transportation operation.
- *Quantitative Risk Assessment of Hazardous Materials Transport Systems* (1996), M. Nicolet-Monnier and A.V. Gheorge, Kluwer Academic Publishers. This book contains a comprehensive treatment of the analysis and assessment of transport risks due to hazmat transport on roads, rail, by ship, and pipeline. It contains European case studies as well as a discussion of computer-based DSS (Decision Support System) for hazmat transport problems. It is a useful reference book in the area.

2.3 Reports

In addition to these books, there are also a number of recent government reports that contain a wealth of useful information for researchers in OR as well as other relevant fields:

- *AND-DIS-01-1 A National Risk Assessment for Selected Hazardous Materials in Transportation* (2000), Argonne National Laboratory, Department of Energy.
- *ANL-DIS-00-1: Development of the Table of Initial Isolation and Protective Action Distances for the 2000 Emergency Response Guidebook* (2000), Argonne National Laboratory, Department of Energy.

- *Comparative Risks of Hazardous Materials and Non-Hazardous Materials Truck Shipment Accidents/Incident (2001)*, Batelle.
- *A Resource Handbook on DOE Transportation Risk Assessment* (2002), DOE Transportation Risk Assessment Working Group Technical Subcommittee.

(Note: All URLs in this chapter were functional as of May 2005.)

2.4 Web sites

The following web sites contain useful information for practitioners as well as researchers on hazmat transport:

- The Office of Hazardous Materials Safety (US DOT Research and Special Programs Administration): *http://hazmat.dot.gov/*.
- The Hazmat 101 Web: *http://www.hazmat101.com/*.
- Hazmat Magazine: *http://www.hazmatmag.com/*.
- On-line hazmat school: *http://www.hazmatschool.com/*.
- National Hazardous Materials Route Registry: *http://hazmat.fmcsa.dot.gov/*.
- United Nations Economic Commission for Europe (UNECE) – Dangerous Goods and Special Cargo: *http://www.unece.org/trans/danger/danger.htm*.
- The Canadian Transport Emergency Centre (CANUTEC) of the Department of Transport: *http://www.tc.gc.ca/canutec/*.
- A mailing list for those interested in hazmat transport: *http://groups.yahoo.com/group/DangerousGoods/*.

2.5 Software

There exists some software which has been developed to aid the analysts or decision makers in dealing with hazmat logistics. For example, ALOHA (Areal Locations of Hazardous Atmospheres) predicts how a hazardous gas cloud might disperse in the atmosphere after an accidental chemical release. This software (see US EPA, 2004) has been developed jointly by the National Oceanic and Atmospheric Administration's (NOAA) Hazardous Materials Response and Assessment Division and the US Environmental Protection Agency's (EPA) Chemical Emergency Preparedness and Prevention Office. ALOHA can be useful for transport risk assessment. However, this software is more useful for fixed facility risk assessment than for route selection.

In contrast to the availability of many software packages for regular truck routing, we know of only one off-the-shelf hazmat routing package that is currently available: PC*Miler|HazMat (ALK Associates, 1994). It has features that allow transportation and logistics companies to determine routes and mileages for hauling hazardous materials while ensuring compliance with

government regulations. Routes can be generated for general, explosive, inhalant and radioactive hazmats. This software contains all of the features and functionality of PC*Miler, a routing, mileage and mapping software, which is also developed by ALK. Here we note that HazTrans (Abkowitz et al., 1992) and PC*HazRoute (ALK Associates, 1994) were marketed in the last decade, but both are off the market as of 2005.

2.6 *Classification*

While we offer references to books, reports, and web sites in this section, the rest of this chapter deals mainly with the academic literature consisting of refereed journal articles. Figure 3 displays the number of papers published in this area between 1982 and 2004. It seems that this area of research has peaked in mid-1990s and has declined somewhat since.

Given the large number of papers in this area, we believe a simple classification can be useful in providing some structure to the rest of the chapter. The articles in this area deal with different aspects of the problem. One possible classification is the following (in no particular order):

(1) risk assessment,
(2) routing,
(3) combined facility location and routing,
(4) network design.

Although we have offered this simple classification, it is fair to say that numerous papers deal with problems that lie at the intersection of the above areas and such problems are receiving increasingly more attention in the literature.

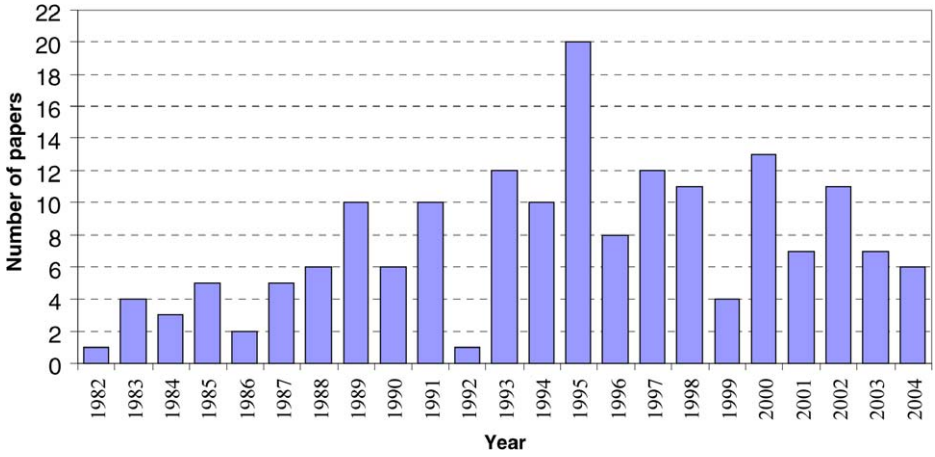


Fig. 3. Number of hazmat-transportation related papers published in refereed journals between 1982 and 2004.

Tables 2(a–d) provides a classification of papers using the above four problem classes as well as other important attributes such as transport mode, paradigm (deterministic vs. stochastic) and number of objectives, and whether or not the paper uses GIS (Geographic Information System) or proposes a DSS.

The rest of this chapter provides a comprehensive literature survey following the problem classification presented above, and points out directions for future research.

Table 2a.
A classification of hazmat transportation models – risk assesment

Road	Jonkman et al., 2003; Nardini et al., 2003; Martinez-Alegria et al., 2003[G]; Rosmuller and Van Gelder, 2003; Abkowitz, 2002[C]; Fabiano et al., 2002; Kimberly and Killmer, 2002; Saccomanno and Haastrup, 2002[N]; Hollister, 2002; Hwang et al., 2001; Abkowitz et al., 2001; Verter and Kara, 2001[G]; Efroymson and Murphy, 2000; ICF Consulting, 2000; Leonelli et al., 2000; Zhang et al., 2000[G]; Pet-Armacost et al., 1999; Cassini, 1998; Mills and Neuhauser, 1998; Cutter and Ji, 1997; Groothuis and Miller, 1997; Lovett et al., 1997[G]; Pine and Marx, 1997; Alp and Zelensky, 1996; Ertugrul, 1995; Sissell, 1995; Chakraborty and Armstrong, 1995; Erkut and Verter, 1995a[U]; Erkut and Verter, 1995b; Moore et al., 1995[G]; Spadoni et al., 1995; Verter and Erkut, 1995[U]; Gregory and Lichtenstein, 1994; Macgregor et al., 1994; Hobeika and Kim, 1993; Sandquist et al., 1993; Harwood et al., 1993; Abkowitz et al., 1992; Glickman, 1991; Grenney et al., 1990[DSS]; Kunreuther and Easterling, 1990; Chow et al., 1990; Abkowitz and Cheng, 1989; Ang and Briscoe, 1989; Harwood et al., 1989; Abkowitz and Cheng, 1988; Hillsman, 1988; Horman, 1987; Keeney and Winkler, 1985; Scanlon and Cantilli, 1985; Pijawka et al., 1985; Kunreuther et al., 1984; Philipson et al., 1983; Wilmot, 1983; Keeney, 1980; Shappert et al., 1973
Rail	Anderson and Barkan, 2004; Barkan et al., 2003; Fronczak, 2001; Orr et al., 2001; Dennis, 1996; Larson, 1996; Glickman and Golding, 1991; McNeil and Oh, 1991; Saccomanno and Elhage, 1991; Glickman and Rosenfield, 1984; Glickman, 1983; Saccomanno and El-Hage, 1989
Marine	Douligeris et al., 1997; Roeleven et al., 1995; Romer et al., 1995
Air	LaFrance-Linden et al., 2001
Road + rail	Brown and Dunn, 2007; Milazzo et al., 2002; Bubbico et al., 2000; Neill and Neill, 2000; Deng et al., 1996; Leeming and Saccomanno, 1994; Purdy, 1993; Saccomanno and Shortreed, 1993; Saccomanno and El-Hage, 1989; Vanaerde et al., 1989; Glickman, 1988; Swoveland, 1987
Road + rail + marine	Andersson, 1994
Road + rail + marine + air	Kloeber et al., 1979

[C]with security consideration;
[DSS]decision support system model;
[G]using GIS;
[N]through road tunnels;
[U]survey/annotated bibliography.

Table 2b.
A classification of hazmat transportation models – routing

Local routing	Road	Akgün et al., 2007; Duque, 2007; Erkut and Ingolfsson, 2005; Huang and Cheu, 2004[CG]; Huang et al., 2003[CM]; Kara et al., 2003; Luedtke and White, 2002[CU]; Marianov et al., 2002; Frank et al., 2000; Erkut and Ingolfsson, 2000; Leonelli et al., 2000; Zografos et al., 2000[DSS]; Erkut and Verter, 1998; Tayi et al., 1999[M]; Bonvicini et al., 1998; Marianov and ReVelle, 1998[M]; Verter and Erkut, 1997; Sherali et al., 1997[M]; Nembhard and White, 1997[M]; Erkut and Glickman, 1997; Jin and Batta, 1997; Verter and Erkut, 1997; Erkut, 1996; Jin et al., 1996; Ashtakala and Eno, 1996[S]; Beroggi and Wallace, 1995; Boffey and Karkazis, 1995; Erkut, 1995; Moore et al., 1995[G]; Karkazis and Boffey, 1995; Glickman and Sontag, 1995[M]; McCord and Leu, 1995[M]; Sivakumar et al., 1995; Beroggi, 1994; Beroggi and Wallace, 1994; Ferrada and Michelhaugh, 1994; Patel and Horowitz, 1994[G]; Sivakumar and Batta, 1994; Lassarre et al., 1993[G]; Sivakumar et al., 1993; Turnquist, 1993[MS]; Wijeratne et al., 1993[M]; Lepofsky et al., 1993[G]; Beroggi and Wallace, 1991; Miaou and Chin, 1991; Gopalan et al., 1990a; Chin, 1989[M]; Zografos and Davis, 1989[M]; Abkowitz and Cheng, 1988[M]; Batta and Chiu, 1988; Vansteen, 1987; Cox and Turnquist, 1986; Belardo et al., 1985; Saccomanno and Chan, 1985; Urbanek and Barber, 1980; Kalelkar and Brinks, 1978[M]
	Rail	Verma and Verter, 2007; McClure et al., 1988; Coleman, 1984; Glickman, 1983
	Marine	Iakovou, 2001; Li et al., 1996; Haas and Kichner, 1987
	Road + rail	Glickman, 1988
	Road + rail + marine	Weigkricht and Fedra, 1995[DSS]
Local routing and scheduling (on road)		Erkut and Alp, 2006; Chang et al., 2005[MST]; Zografos and Androutsopoulos, 2004[M]; Zografos and Androutsopoulos, 2002[M]; Miller-Hooks and Mahmassani, 2000[ST]; Bowler and Mahmassani, 1998[T]; (Miller-Hooks and Mahmassani, 1998)[ST]; Sulijoadikusumo and Nozick, 1998[MT]; (Nozick et al., 1997)[MT]; Smith, 1987[M]; Cox and Turnquist, 1986
Global routing	Road	Carotenuto, et al. (2007a, 2007b); Dell'Olmo et al., 2005; Akgün et al., 2000; Marianov and ReVelle, 1998; Lindner-Dutton et al., 1991; Gopalan et al. (1990a, 1990b); Zografos and Davis, 1989
	Marine	Iakovou et al., 1999

[C]with security consideration;
[DSS]decision support system model;
[G]using GIS;
[M]multiobjective;
[S]stochastic;
[T]time-varying;
[U]survey/annotated bibliography.

Table 2c.
A classification of hazmat transportation models – combined facility location and routing

Alumur and Kara, 2007; Cappanera et al., 2004; Berman et al., 2000; Giannikos, 1998[M]; Helander and Melachrinoudis, 1997; List and Turnquist, 1998; Current and Ratick, 1995[M]; Jacobs and Warmerdam, 1994; Boffey and Karkazis, 1993; Stowers and Palekar, 1993; List and Mirchandani, 1991[M]; List et al., 1991[U]; ReVelle et al., 1991; Zografos and Samara, 1989; Peirce and Davidson, 1982; Shobrys, 1981

[M]multiobjective;
[U]survey/annotated bibliography.

Table 2d.
A classification of hazmat transportation models – network design

Berman et al., 2007; Erkut and Alp, 2006; Erkut and Gzara, 2005; Erkut and Ingolfsson, 2005; Verter and Kara, 2005; Kara and Verter, 2004

3 Risk assessment

Risk is the primary ingredient that separates hazmat transportation problems from other transportation problems. In this section we will provide a detailed treatment of how risk is incorporated into hazmat transport models, starting with the basic building blocks and moving our way into risk assessment along a route. In the context of hazmat transport, risk is a measure of the probability and severity of harm to an exposed receptor due to potential undesired events involving a hazmat (Alp, 1995). The exposed receptor can be a person, the environment, or properties in the vicinity. The undesired event in this context is the release of a hazmat due to a transport accident. The consequence of a hazmat release can be a health effect (death, injury, or long-term effects due to exposure), property loss, an environmental effect (such as soil contamination or health impacts on flora and fauna), an evacuation of nearby population in anticipation of imminent danger, or stoppage of traffic along the impacted route.

Risk assessment can be qualitative or quantitative. Qualitative risk assessment deals with the identification of possible accident scenarios and attempts to estimate the undesirable consequences. It is usually necessitated by a lack of reliable data to estimate accident probabilities and consequence measures. The goal is to identify events that appear to be most likely and those with the most severe consequences, and focus on them for further analysis. It may be the only option in the absence of data – for example, assessing the risks due to the location of a permanent nuclear waste repository. While hazmat transport analysts are known to complain about the quality of their data (we will return to this topic later in this section), they do have access to considerable historical information on accident frequencies and fairly accurate consequence models for hazmat releases in case of accidents in many developed countries.

Due to this, and the necessity of quantitative information for OR models, in this section we focus on quantitative risk assessment.

Quantitative risk assessment (QRA) involves the following key steps:

(1) hazard and exposed receptor identification;
(2) frequency analysis; and
(3) consequence modeling and risk calculation.

Identification of hazard refers to identifying the potential sources of release of contaminants into the environment, the types (e.g., thermal radiation due to jet and pool fires and fireballs, explosions, flying pieces of metals or other objects due to blast waves, toxic clouds, and flame) and quantities of compounds that are emitted or released, and the potential health and safety effects associated with each substance. In some cases (for example, when a release of carcinogenic substances is involved), we also need to investigate the long-term health risks of a hazmat accident. Examination of risks on different types of exposed receptor is also essential to cover different response characteristics in the risk assessment.

The language of QRA is one of *frequencies* and *consequences*, and unlike in qualitative risk analysis, QRA results in a numerical assessment of risks involved, for example, an expected number of individuals impacted per year. In the next two sections we discuss frequency analysis and consequence modeling along with risk calculation.

3.1 Frequency analysis

The frequency analysis involves (a) determining the probability of an undesirable event; (b) determining the level of potential receptor exposure, given the nature of the event; and (c) estimating the degree of severity, given the level of exposure (Ang and Briscoe, 1989). Each stage of this assessment requires the calculation of a probability distribution, with stage (b) and (c) involving conditional distributions. Consider a unit road segment. Suppose that there is only one type of accident, release, incident, and consequence. Let A be the accident event that involves a hazmat transporter, M the release event, and I the incident event. Suppose that the consequence of the hazmat release is expressed in terms of the number of injuries. We denote the event of an injury to an individual as D. Using *Bayes' theorem*, we obtain the probability of an injury resulting from an accident related to the hazmat as

$$\begin{aligned} p(A, M, I, D) &= p(D|A, M, I)\, p(A, M, I) \\ &= p(D|A, M, I)\, p(I|A, M)\, p(A, M) \\ &= p(D|A, M, I)\, p(I|A, M)\, p(M|A)\, p(A), \end{aligned} \tag{3.1}$$

where $p(E)$ denotes the probability of the event E occurring on the road segment and $p(E|F)$ the associated conditional probability. Despite its simplicity, the above model already contains many of the necessary elements for hazmat

risk assessment. For example, Chow et al. (1990) used a Bayesian model that includes multiple levels of event severity to predict severe nuclear accidents and to estimate the associate risks. Glickman (1991) used a Bayesian model in the assessment of the risks of highway transportation of flammable liquid chemicals in bulk.

Furthermore, let s_{lm} denote the number of shipments of hazmat m on road segment l per year. Note that a highway transport route from the origin to the destination consists of finitely many road segments. The product $s_{lm}\, p_l(A, M_m, I, D)$ determines the frequency of the occurrence of the hazardous release event that measures the *individual risk* for a person in the neighborhood of road segment l. Usually, the individual risk is defined as the yearly death frequency for an average individual at a certain distance from the impact area (see, e.g., Mumpower, 1986; Leonelli et al., 2000). Although no universally accepted individual risk criteria exist, one tends to compare the risk of death to de minimis of 10^{-6} to 10^{-5} deaths per year (Mumpower, 1986).

Hazmat incidents usually impact a number of individuals. Hence, we need to move from individual risk toward *societal risk*. The societal risk is a characteristic of the hazardous activity in combination with its populated surroundings. There are several ways to express societal risk. Perhaps the simplest method is to compute the expected number of impacted individuals by multiplying the probability of impact per person with the number of persons present in the impact zone. Hence, the societal risk (or just *risk* for short) on road segment l of hazmat m, R_{lm}, can be expressed as

$$R_{lm} := s_{lm} \iint_L p_l(D_{xy}|A, M_m, I)\, p_l(I|A, M_m) \times p_l(M_m|A)\, p_l(A) POP_l(x, y)\, \mathrm{dx}\, \mathrm{dy}, \tag{3.2}$$

where $p_l(D_{xy}|A, M_m, I)$ is the probability that individuals on location (x, y) in the impact area L will be dead due to the incident on a route segment l and $POP_l(x, y)$ is the population density on location (x, y) in the neighborhood of road segment l. By assuming that each individual in the affected population will incur the same risk, R_{lm} can be simply expressed as

$$R_{lm} := s_{lm}\, p_l(D|A, M_m, I)\, p_l(I|A, M_m)\, p_l(M_m|A)\, p_l(A) POP_l. \tag{3.3}$$

Thus, if few people are present around the hazardous activity, the societal risk may be close to zero, whereas the individual risk may be quite high.

While this expected consequence is a convenient measure for OR models, the risk assessment literature prefers a richer measure, namely the FN-curve which expands the point estimate of the expectation to the entire distribution. To produce an FN-curve, one has to compute the probability that a group of more than N persons would be impacted due to a hazmat accident, for all levels of N. The risk level is communicated by the FN-curve, a graph with the ordinate representing the cumulative frequency distribution F of the hazardous release events which result in at least N number of impacts (e.g.,

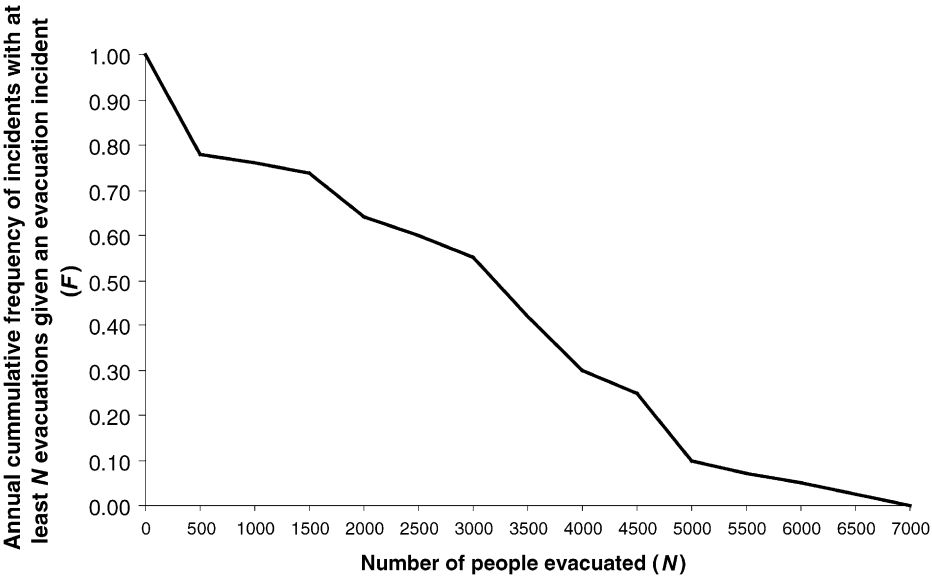


Fig. 4. A conditional FN-curve (given an evacuation incident).

number of fatalities or number of people evacuated) and abscissa representing the consequence (N impacts). Furthermore, if F is a conditional cumulative frequency distribution, then the associated FN-curve is called the *conditional* FN-curve. Figure 4 shows a conditional FN-curve for PCB transport through Edmonton, Canada (Erkut and Verter, 1995b). The ordinate F is the annual cumulative frequency of incidents with at least N evacuations conditioned on the occurrence of an evacuation incident in the city. This figure shows that if an evacuation incident occurs, then the probability of evacuating more than 500 people is 0.8. Some countries (such as Denmark, Netherlands and the UK) use decision rules for hazmat installations based FN-curves (Jonkman et al., 2003).

Clearly, more than one type of accident, release, incident, and consequence can occur during the hazmat transport activity. For example, a release of flammable liquid can lead to a variety of incidents such as a spill, a fire, or an explosion. To accommodate this, let $\mathcal{A}$, $\mathcal{M}$, $\mathcal{I}$, and $\mathcal{C}$ denote the set of possible accidents, releases, incidents, and consequences that may occur on road segment l. Suppose that all consequences (injuries and fatalities, property damage, and environmental damage) can be expressed in monetary terms (see Section 3.2.3). Then, the hazmat transport risk associated with road segment l can be expressed as

$$R_l := \sum_{a\in\mathcal{A}} \sum_{m\in\mathcal{M}} \sum_{i\in\mathcal{I}} \sum_{c\in\mathcal{C}} s_{lm} p_l(A_a, M_m, I_i, C_c) CONS_c, \tag{3.4}$$

where $CONS_c$ is the possible c-type consequence.

To summarize, we started with individual risk due to a single incident, then we moved on to risk due to multiple shipments, and on to societal risk, and finally to societal risk with multiple incidents.

However, in practice, researchers frequently neglect these conditional probabilities and simplify the analysis by considering the *expected loss* (or the *worst-case loss*) as the measure of risk. The expected value is calculated as the product of the probability of a release accident and the consequence of the incident (List et al., 1991). Hence the hazmat transport risk associated with a road segment l can be expressed as

$$R_l := \sum_{m \in \mathcal{M}} s_{lm} p(M_m) c_{lm}, \tag{3.5}$$

where c_{lm} is the undesirable consequence due to the release of hazmat m on road segment l. This risk model is sometimes referred to as the *technical risk* (Erkut and Verter, 1998). The US DOT use this expected loss definition in their guidelines for transporting hazmats (US DOT, 1994). These simplifications are mainly due to the lack and inaccuracy of accident and exposure data.

As it is clear from the discussion above, QRA depends heavily on an estimation of probabilities. There are two primary means to estimate the accident, release, and incident probabilities: historical frequencies and logical diagrams (fault tree and event tree analysis).

Historical frequencies

We can use the number of hazmat transport accidents in a given time period and the total distance traveled by hazmat trucks in the same time period to calculate the accident rate on a unit road segment (i.e., accidents per km). The hazmat accident probability on road segment l, $p_l(A)$, can be obtained by multiplying the accident rate by the length of road segment l. To estimate $p_l(M_m|A)$, we need to calculate the percentage of hazmat accidents that result in a release of hazmat m. Similarly, we can use historical data to estimate $p_l(I|A, M_m)$ and $p_l(D|A, M_m, I)$. However, the occurrence of an accident may be influenced by intrinsic factors such as tunnels, rail bridges, road geometry, weather conditions, and human factors, as well as other factors correlated to traffic conditions, such as traffic volume and frequency of hazmat shipment. Consequently, some locations are more vulnerable to accidents than others. Therefore, a careful analysis should be done prior to the use of historical data. The rarity of hazmat accidents may result in insufficient information to determine whether historical figures are relevant to the circumstances of concern, particularly regarding rare catastrophic accidents. Moreover, in estimating the associate probabilities on road segment l of a hazmat transportation route, the dependency to the impedances of preceding road segments should also be taken into account (Kara et al., 2003; Verter and Kara, 2001). We will discuss this dependency issue in more detail in Section 3.3.1.

Logical diagram-based techniques

An alternative way to estimate the frequency (and possibly consequences) of hazmat release incidents is the use of logical diagram-based techniques, namely fault tree and event tree analysis. *Fault Tree Analysis* (FTA) is a top-down analysis tool to identify the causes of events and to quantify various accident scenarios that would cause the system fail. It starts with an identified hazard (e.g., chlorine release due to a transport accident) as the root of a tree (or top event) and works backward to determine its possible causes (e.g., collision accident, derailment, and relief valve poorly sealed) using two logical functions: OR and AND. The causative events are laid out in a tree with the branches connected by gates comprising one of these logical functions. The OR gate represents the union of events attached to the gate. An OR gate can have any number of inputs (branches). The event above the gate is realized if any one or more of the inputs occur. The AND gate represents the intersection of events attached to the gate. An AND gate can also have any number of inputs, but the event above the gate is only realized if all the inputs occur. Moreover, several fault trees can be combined into a single complex fault tree. FTA enables us to determine the probability of the top event on the basis of the probabilities of the basic events (e.g., $p(D|A, M, I)$ in (3.1), where death of an individual in hazmat transport accident is the top event) for which sufficient historical data exist or expert judgments are reliable.

Event Tree Analysis (ETA), on the other hand, is a bottom-up analysis tool. It takes at its starting point the event that can affect the system (e.g., an initial release of hazmat) and tracks them forward through sequences of interfacing system components to determine their possible consequences. It examines all possible responses to the initiating event, such as the functioning, failure, or partial failure of subsystems or different systems, in a tree structure with the branches developing from left to right. Each outcome of the branches is usually binary (i.e., the outcome occurs or does not occur). By assigning a probability to each branch, the probabilities of every possible outcome following the initiating event can be determined. ETA can be used in conjunction with FTA, called FETA, to identify and quantify the possible consequences of the top event in fault tree. Figure 5 shows a fault tree in conjunction with an event tree. For additional details and examples of fault and event tree construction, we refer to Henley and Kumamoto (1981), Vesely et al. (1981), and Rhyne (1994).

Boykin et al. (1984) applied FETA to analyze the risks associate with the selection of technology alternatives in the chemical storage system. Pet-Armacost et al. (1999) used FETA in conjunction with two Monte Carlo simulations (one uses spreadsheet add-in @RISK and the other uses discrete event simulation software ARENA) to conduct a transportation risk analysis of Hydrazine in order to determine whether or not a relief valve should be used. FETA was used to decompose the transport process into its basic components and to identify the major sources of uncertainty. The event probabilities in the event trees were derived as functions of the parameters whose effects were

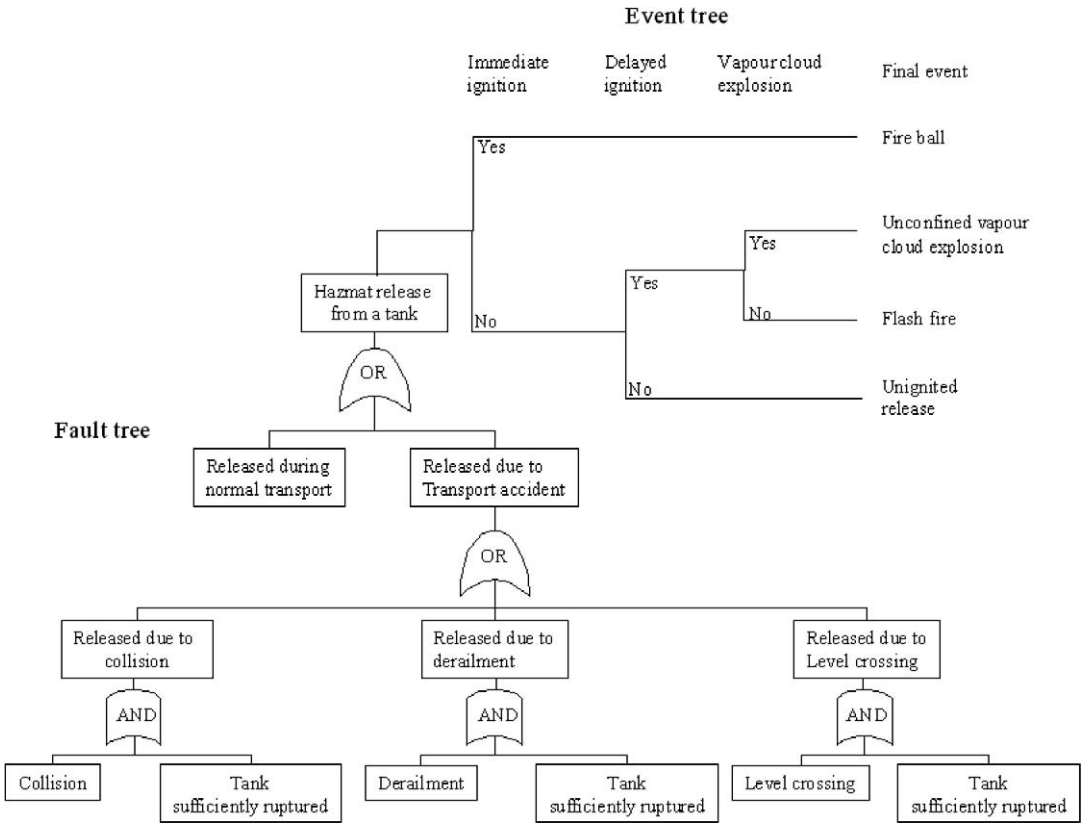


Fig. 5. A fault tree in conjunction with an event tree for hazmat release (adapted from Alp, 1995).

not known. The impact of these parameters on the risks of toxic exposure, fire, and explosion was analyzed through Monte Carlo analysis and analysis of variance. Rosmuller and Van Gelder (2003) used FETA to conduct a QRA for the hazmat transportation in the Netherlands. The results were used to formulate appropriate risk and rescue policies. They suggested that emergency response teams could use the release data for determining impact circles for road accidents and subsequently decide on detour routes. Moreover, expected distributions of release quantities could be used to facilitate the training of hazmat response personnel.

3.2 Consequence modeling and risk calculation

3.2.1 Modeling the impact area

There are many undesirable consequences of a hazmat transportation accident, such as economic losses, injuries, environmental pollution, damage to wildlife, and fatalities. These consequences are a function of the impact area (or exposure zone) and population, property, and environmental assets within the impact area. The shape and size of an impact area depends not only on the substance being transported but also on other factors, such as topology, weather, and wind speed and direction. Estimating, a priori, the impact area of a potential accident is difficult. Researchers used different geometric shapes to model the impact area such as a band of fixed width around each route segment (e.g., Batta and Chiu, 1988; ReVelle et al., 1991), a circle (called danger circle), with a substance-dependent radius centered at the incident location (e.g., Erkut and Verter, 1998; Kara et al., 2003), rectangle around the route segment (e.g., ALK Associates, 1994), and an ellipse shape based on the Gaussian plume model (e.g., Patel and Horowitz, 1994; Chakraborty and Armstrong, 1995; Zhang et al., 2000). Figure 6 shows these four shapes of the impact area that have been used in the literature.

Perhaps the most common approximation of the impact zone is the *danger circle*. By moving the danger circle along a route segment between two nodes (see Kara et al., 2003), we get the fixed-bandwidth approximation and by cutting off the circular segments at the two ends we get the rectangle approximation. The bandwidth or radius is substance-dependent but it is assumed to be constant for a given shipment, which means that this approximation does not consider effect of the distance on the level of impact. One can determine the radius by considering the *evacuation distance* (i.e., the *initial isolation zone*) when a hazmat incident occurs, for example, 0.8 km for flammable hazmats and 1.6 km for flammable and explosive hazmats (CANUTEC, 2004). The central assumption in these models is that each individual within the danger zone will be impacted equally and no one outside of this area will be impacted.

The modeling of an impact area can also be considered from the point of view of the affected population center. For example, a population center is commonly modeled as a point on the plane, where all inhabitants of the

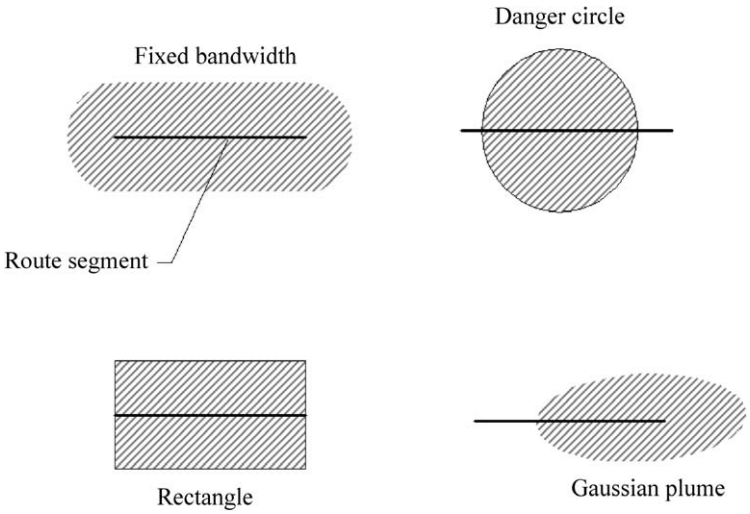


Fig. 6. Possible shapes of impact area around the route segment.

population center are considered to experience the same impact from a hazmat incident on a road segment nearby. The impact on this aggregation point depends on the distance between the point and the incident location. For example, the impact can be inversely proportional to the square of the Euclidean distance between the two points. However, a GIS enables researchers to represent the spatial distribution of population density more accurately (see, for example, Figure 8) rather than using aggregation points. Erkut and Verter (1995b) proposed a model of the spatial distribution of population by using a polygon. Verter and Kara (2001) incorporated this in a GIS, and developed a large-scale risk assessment model for the provinces of Ontario and Quebec.

3.2.2 Gaussian plume model

In an airborne hazmat (e.g., chlorine, propane, and ammonia) accident, the concentration of the airborne contaminant varies with distance from the source of accident. It will be lower as the gas disperses with distance and wind. Therefore, the three approaches discussed above can result in poor approximations of the impact area. In this case, researchers have usually resorted to the Gaussian plume model (GPM) (Hanna et al., 1993; Patel and Horowitz, 1994; Chang et al., 1997; Zhang et al., 2000; Puliafito et al., 2003). The Gaussian plume model is based on several limiting assumptions:

(1) the gas does not change its chemical properties during dispersion;
(2) the terrain is unobstructed and flat;
(3) the ground surface does not absorb the gas;
(4) the wind speed and direction is stable during the dispersion period; and
(5) the emission rate is constant.

These assumptions certainly limit the application of GPM, for example, assumption (1) restricts the applicability of the GPM to stable chemicals and to accidents which do not result in an explosion (Zhang et al., 2000). The GPM is formulated as

$$C(x, y, z, h_e) = \frac{Q}{2\pi\mu\sigma_y\sigma_z} \exp\left(-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right) \times \left[\exp\left(-\frac{1}{2}\left(\frac{z-h_e}{\sigma_z}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\frac{z+h_e}{\sigma_z}\right)^2\right)\right],$$

where C is the concentration level (mass per unit volume – $\mu g/m^3$ or parts per million – ppm), x is the distance downwind from the source (m), y is the distance crosswind (perpendicular) from the source (m), z is the elevation of the destination point (m), h_e is the elevation of the source (m), Q is the release rate of pollutant (mass emission rate – g/s or volumetric volume rate – m^3/s), μ is the average wind speed (m/s), σ_y and σ_z are the dispersion parameters in the y and z directions (m).

In hazmat dispersion from traffic accidents, it is usually assumed that the source is on the ground (i.e., $h_e = 0$) and we are interested in the ground concentration level (i.e., $z = 0$). Therefore, we obtain

$$C(x, y, z, h_e) = C(x, y) = \frac{Q}{\pi\mu\sigma_y\sigma_z} \exp\left(-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right).$$

Figure 7 shows bell-shaped curves of concentration levels $C(x, y)$ for two different downwind distances: (a) the concentration of the pollutant is high at the source of the spill ($x = 0$) and the Gaussian distribution has a pronounced peak; (b) as the pollutant drifts farther downwind ($x \gg 0$), it spreads out and the bell-shape becomes wider and flatter.

The release rate, Q, depends on container volume, hazmat type, and rupture diameter. To calculate Q, one can use ALOHA (see Section 2.5). ALOHA can also be used for estimating the concentration level, $C(x, y)$, but its results are only reliable within one hour of the release event, and 10 kilometers from the release source. The dispersion parameters, σ_y and σ_z, can be determined as a function of downwind distance x (Pasquill and Smith, 1983; Arya, 1999).

The individual risk, that is the probability that an individual at location j with coordinate (j^x, j^y) will experience an undesirable consequence (such as evacuation, or injury, or death) as a result of a release at i, p_{ij}, can be represented as a function of the concentration of airborne contaminant at j, $C_{ij} := C(|j^x - i^x|, |j^y - i^y|)$. The American Institute of Chemical Engineers (2000) suggests a *probit function* to model $p_{ij}(C_{ij})$. Consequently, the social risk can be obtained by multiplying $p_{ij}(C_{ij})$ with the population size at location j.

A simpler alternative way to estimate the consequence of airborne hazmat accident is to use the standard concentration level to determine the threshold

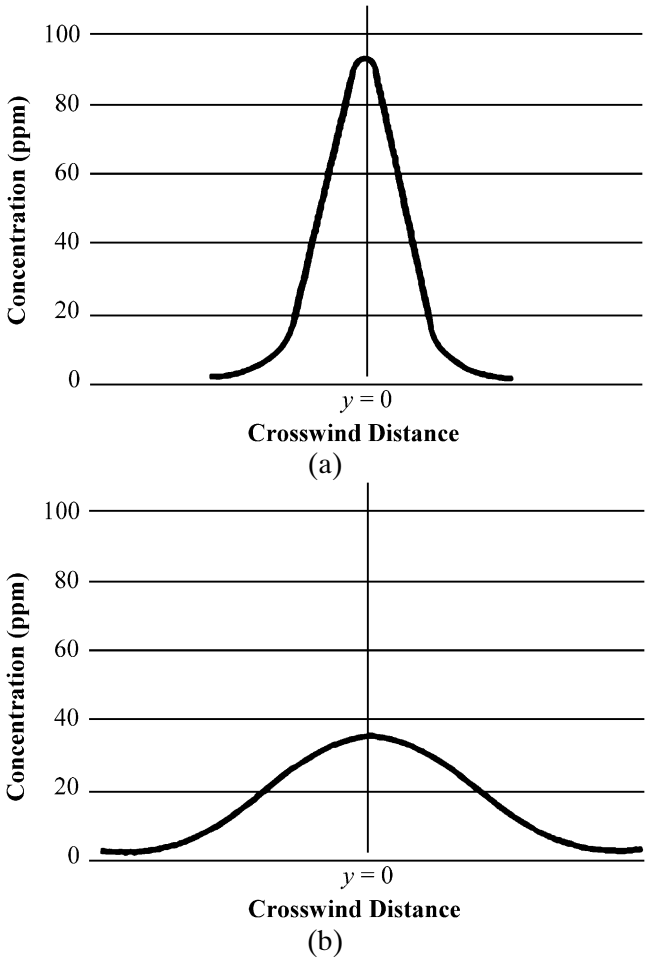


Fig. 7. The bell-shape of concentration level $C(x, y)$: (a) Gaussian distribution at $x = 0$ and (b) Gaussian distribution at $x \gg 0$ (Chakraborty and Armstrong, 1995).

distances for different consequences (e.g., fatalities and injuries), such as Immediately Dangerous to Life and Health (IDLH) (NIOSH, 1994) developed by the National Institute for Occupational Safety and Health (NIOSH) and the Occupational Safety and Health Administration (OSHA). The IDLH-values represent the maximum concentration from which one could escape without injury or irreversible health effects (e.g., severe eye or respiratory irritation, disorientation, or lack of coordination) within 30 minutes of exposure. For example, the IDLH-values for carbon dioxide and propane are 40,000 ppm and 2100 ppm respectively. These numbers hold for enclosed spaces (and not open-air). To be used in an open environment, for example, Verma and Verter (2007) considered a propane dispersion of 2100 ppm per second and assumed that

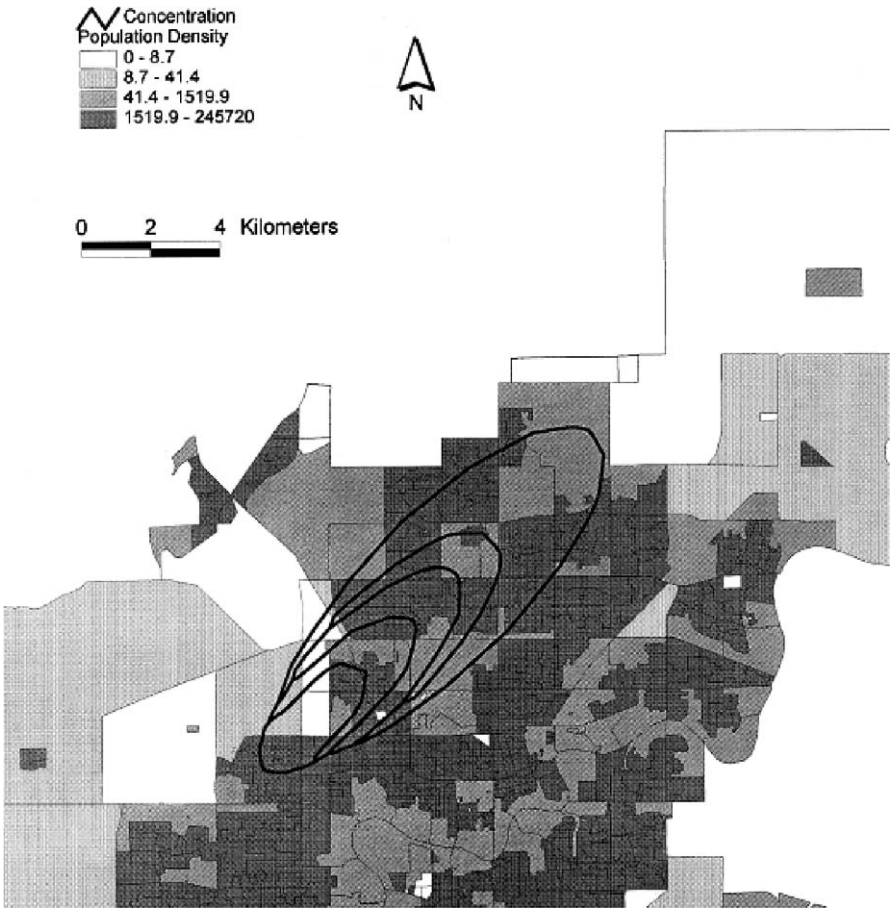


Fig. 8. Population densities within different concentration levels (Zhang et al., 2000).

roughly a 4–5 minute propane exposure at this IDLH level can cause minor injury while a 30–35 minute exposure can cause major injury or fatality. Using these assumptions, they defined a fatality zone (if the concentration level $C \geqslant 4{,}200{,}000$), an injury zone (if $600{,}000 \leqslant C < 4{,}200{,}000$) and a nonexposure zone (if $C < 600{,}000$) where C is given in ppm. Hence, the threshold distance is determined by the level curve of the associated hazmat IDLH-value and the associated consequence can be represented as the function of the population size within the level curve. Figure 8 shows the population densities within different concentration levels of a single source release.

The following conceptual example demonstrates how an improper assessment of the impact area may lead to a high-risk routing decision. Consider two east–west routes around a city that may be used for propane shipments: South (P_1) and North (P_2) routes (see Figure 9(a)). Assume each route seg-

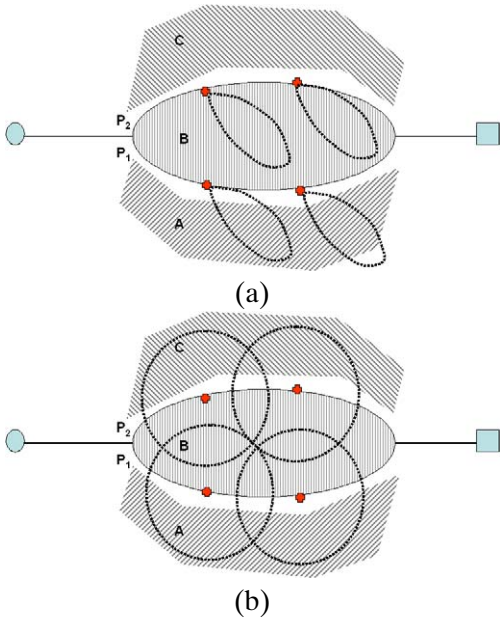


Fig. 9. (a) Gaussian plume model vs. (b) danger circle.

ment in both routes has the same incident probability. Suppose these routes divide the city into three regions A, B, C, where each region has uniform population density. Among these regions, suppose that region B is the most densely populated one and region C is the least densely populated one. Moreover, suppose that the prevalent wind direction is south-east. Figure 9(b) shows concentration contours of route segments in P_1 and P_2, according to the IDLH value. Since the population density in the impact area of route P_1 is less than that of P_2, one might send propane via route P_1. In contrast, if one were to use a danger circle instead of the Gaussian plume model, neglecting the type of hazmat and the wind direction, one may select route P_2 instead of P_1. This decision would expose more people in case of an incident as propane would drift south-eastwardly into region B. As this simple example demonstrates a careful analysis is necessary prior to defining the impact area.

3.2.3 Risk cost

To estimate the cost of a hazmat release incident, various consequences must be considered. The consequences can be categorized into (Abkowitz et al., 2001; FMCSA, 2001): injuries and fatalities (or often referred to as *population exposure*), cleanup costs, property damage, evacuation, product loss, traffic incident delay, and environmental damage. All impacts must be converted to the same unit (for example dollars) to permit comparison and complication of the total impact cost. The discussion of risk costs presented here deals primarily with hazmat incidents on highways.

Injuries and fatalities. Finding a dollar value of human life and safety is perhaps the most difficult and controversial issue. Some find it offensive; others argue that any dollar value assigned to human life would be too low. Yet it is possible to estimate the value indirectly. Insurance payments offer a simple estimate. Perhaps more relevant is the figure used by government agencies to prioritize their projects that reduce fatalities and injuries. Clearly if an agency is making a choice between Project A which will save X lives and cost P dollars per year and Project B which will save Y lives and cost Q dollar per year, they are implicitly using a trade-off value that converts fatalities to dollars – regardless of whether or not the trade-off is made explicit.

The value of an injury or fatality in a hazmat incident can be estimated from different perspectives (FMCSA, 2001). For example, one can value an injury or fatality in terms of lost income and economic productivity to society. The National Highway Transportation Safety Administration (NHTSA) estimates the cost of fatalities and injuries by considering both direct and indirect costs to individuals and to society (NHTSA, 1996). Direct costs include emergency treatment, initial medical costs, rehabilitation costs, long-term care and treatment, insurance administrative expenses, legal costs, and employer/workplace costs. Indirect costs are productivity losses in the workplace due to temporary and permanent disability and decreases in productivity at home resulting from these disabilities. In 1996 dollars, a fatality costs about $913,000 and a critical injury costs about $780,000.

In addition to the economic cost components discussed above, The National Safety Council (NSC) also includes the value of a person's natural desire to live longer or to protect the quality of one's life (NSC, 2003). This value indicates what people are willing to pay to reduce their safety and health risks. Hence, the cost estimates include wage and productivity losses, such as wages and fringe benefits, replacement cost and travel delays caused by the accident; medical expenses, such as doctor fees, hospital charges, cost of medication, future medical costs, and other emergency medical services; administrative expenses, such as insurance premiums and paid claims, police and legal costs; motor vehicle damage, such as property damage to vehicles; and employer costs, such as time lost by uninjured workers, investigation and reporting time, production slowdowns, training of replacement workers, and extra costs of overtime for uninsured workers (FMCSA, 2001). The 2003 estimates of incapacitating injury and fatality costs are $181,000 and $3,610,000, respectively.

Finally, US DOT values injuries and deaths at the amount they would spend to avoid an injury or fatality (FMCSA, 2001). This averages out to be $400,000 to avoid an injury requiring hospitalization and $2,800,000 to avoid a fatality.

Cleanup costs. Cleanup costs are assumed to encompass the costs of both stopping the spread of a spill and removing spilled materials (Abkowitz et al., 2001; FMCSA, 2001). Such costs vary widely depending on the size, type of materials, and location of the spill. Some national database systems, such as the Hazardous Materials Information System (HMIS) of US DOT and The Work-

place Hazardous Materials Information System (WHMIS) of Health Canada, can be used as references for the cleanup costs. For the period 1990–1999, cleanup costs averaged about $24,000 per en-route accident, $1300 per cleanup for an en-route incident spill, and $260 for an unloading/loading accident and incident spill cleanup (HMIS database).

Property damage. Property damage encompasses damage to other vehicles, which may have been involved in the incident, as well as damage to both public and private property (e.g., private buildings, public utilities, public roadways). For example, from HMIS database of the period 1990–1999, the average property damage for flammable and combustible liquids en-route accidents was $16,041, while the average property damage for en-route incident spills was $274. Average property damage for leaks occurring during loading and unloading incidents and accidents was $68. Average property damage for flammable gases en-route accidents, en-route spills, and loading/unloading incidents were $3147, $173, and $2315, respectively. For corrosive materials, the average values for en-route accidents, en-route spill incidents, and loading/unloading incidents were $3104, $67, and $17, respectively (FMCSA, 2001).

Evacuation. There are numerous variables which complicate the estimation of the cost of evacuation. These include the expense for temporary lodging and food, losses due to lost wages and business disruptions, inconvenience to the public, and the cost of agencies assisting in evacuation. A reasonable estimate would be $1000 per person evacuated (TRB, 1993). This $1000 estimate is also used by the Federal Railroad Administration (FRA) to estimate impacts from railroad evacuations.

Product loss. Product loss refers to the quantity and value of the hazmats lost during a spill. For example, from the HMIS database for period 1990–1999, the average cost of product lost of flammable and combustible liquids en-route accident related spills was $3208 per spill. Similarly, for flammable gases accidents, the average cost of product lost per en-route accident related spill was $1140 per spill. Corrosive material spill accidents averaged $4910 per spill in product loss.

Traffic incident delay. Hazmat spills typically require an emergency response that causes a significant traffic delay. This type of traffic delay is called incident delay. If traffic volume and incident situation (e.g., the traffic arrival rate, road capacity reduction, and incident duration) is known, a deterministic model can be used to estimate the incident delay. For example, Morales (1989) used a deterministic queueing model and Wirasinghe (1978) and Alp (1995) used models based on shock-wave theory. Due to its simplicity, the Morales model is often used by practitioners (see, for example, Abkowitz et al., 2001; FMCSA, 2001). However, these deterministic models are inappropriate for prediction of incident delay in real-time situation where the incident duration

is unknown. In this case, incident delay is best modeled by a random variable that represents the stochastic characteristics associated with the incident (as in Fu and Rilett, 1997).

To obtain the associated costs of incident delays, information on the occurrence of an incident or the split between trucks and other vehicles on the various highway systems are required. Earlier studies (Grenzeback et al., 1990) assumed the hourly cost of incident delay to be about $20 for trucks and $10 for other vehicles, which accounts for the value of a driver's time and fuel consumption costs. The total cost traffic incident delay is then obtained by multiplying this dollar value of incident delay with the total number of person-hours of delay given by the model discussed above.

Environmental damage. Environmental damage consists of damage to the environment that remains after the cleanup. This damage can be calculated in terms of loss of economic productivity, such as agricultural production lost and/or in loss of habitat or ecosystem deterioration (FMCSA, 2001). The loss of agricultural productivity can be estimated, for example, using the quantity of crops that are not grown during a 20-year period due to contamination. Using wheat as an example, a contaminated field that can produce 35 bushels per acre/year would result in an (undiscounted) gross income loss of $3500/acre over a 20-year period assuming a fixed value of $5/bushel. To calculate the natural resource environmental damage from a hazmat incident is more complicated. We need to know how much material was spilled, where the spill occurred, and what sort of surface it covered. Using, for example, HMIS data, one can estimate the dollar cost of this damage.

As the discussion in this section points out, while there are different types of costs associated with a hazmat transport incident, in most cases all other costs are dwarfed by the cost of fatalities and injuries and the cost of evacuations in cases of major spills. Perhaps this is a reason why many OR analysts focus exclusively on populations inside a danger circle.

3.2.4 Perceived risks

All consequences we discussed so far assume that society is risk-neutral; i.e., we are indifferent between two consequence distributions, as long as their expected values are equal. For example, risk neutrality assumes that a single incident causing 100 fatalities is equivalent (or equally undesirable to the society) to 100 incidents causing one fatality each, since in both cases the total number of fatalities is the same. However, most individuals would judge a low probability–high consequence (LPHC) event as more undesirable than a high probability–low consequence (HPLC) event even if the expected consequences of the two events are equal (Erkut and Verter, 1998). Consequently when dealing with LPHC events, most human decision makers tend to exhibit risk aversion; a single incident causing 100 fatalities is perceived as much more undesirable than 100 incidents each causing a single fatality.

A simple way to incorporate risk attitude to risk models is to add a risk preference (or tolerance) factor α as an exponent to the consequence values. For example, if the risk assessment deals with the population exposure, then the societal risk on road segment l (see (3.2), dropping the hazmat index m) can be expressed as (see, e.g., Slovic et al., 1984; Abkowitz et al., 1992)

$$R_l := s_l \iint_L p_l(D_{xy}|A, M, I)\, p_l(I|A, M)\, p_l(M|A)\, p_l(A) \times \big(POP_l(x, y)\big)^{\alpha}\, \mathrm{dx}\, \mathrm{dy}.$$

By considering only one shipment (or one trip) and one type of hazmat spill, the traditional expected loss model of risk (3.5) can thus be modified as (see, e.g., Slovic et al., 1984; Abkowitz et al., 1992; Erkut and Verter, 1998; Erkut and Ingolfsson, 2000):

$$R_l := p_l(POP_l)^{\alpha}.$$

Figure 10 shows three different values of α associated with three different risk preferences: $\alpha = 1$ models risk neutrality; $\alpha > 1$ models risk aversion; and $\alpha < 1$ models risk-taking behavior. The greater the value of α, the higher the aversion to the risk of a hazmat incident. The risk-aversion model assumes that the $(i + 1)$st life lost is more costly than the ith life lost, for all possible values of i. Of course as α is increased, any route selection model that operates with an objective of minimizing total risk is eventually reduced to a model that minimizes the maximum risk, as shown by the following small example. Consider a hazmat shipment from an origin O to a destination D. There are two routes (north and south) between O and D, P_1 and P_2, each consisting of two route segments. Suppose that the incident probability and the population density in the impact area of the two segments of route P_1 are $(10^{-4}; 25)$ and $(10^{-4}; 75)$, and those of P_2 are $(10^{-5}; 100)$ and $(10^{-5}; 400)$. The total risks associated with P_1 and P_2 are 10^{-2} and 5×10^{-3}, respectively, and the maximum risks are 75×10^{-4} and 4×10^{-3}, respectively. For $\alpha = 1$, we select P_2, the route with lower total risk. As α approaches infinity, the problem turns into one of minimizing the maximum risk, and we select P_1. Figure 11 shows how

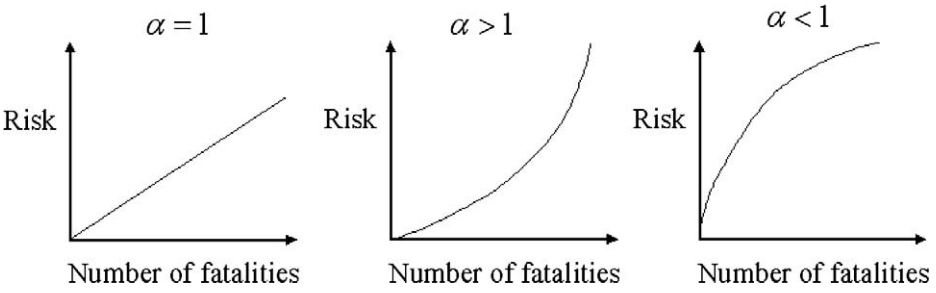


Fig. 10. Three different risk preferences: $\alpha = 1$ risk neutral; $\alpha > 1$ risk aversion; $\alpha < 1$ risk proneness.

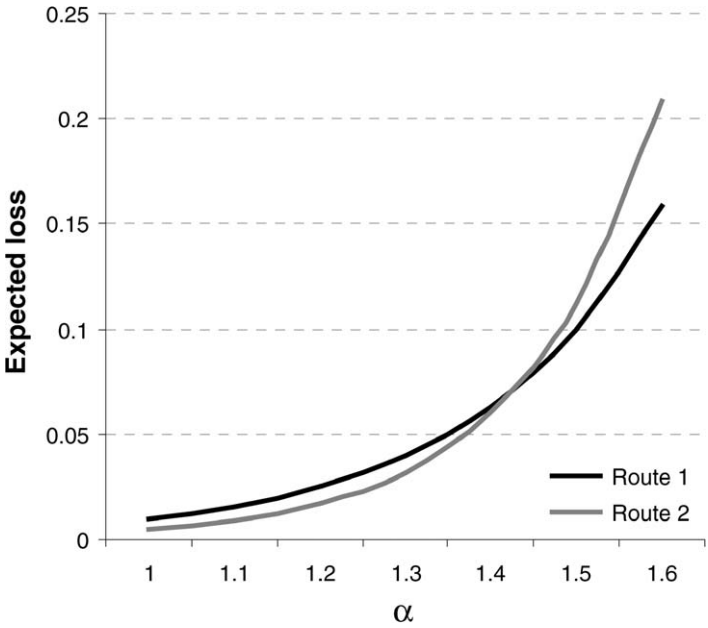


Fig. 11. Routing decision for different values of α, based on perceived risk.

the optimal routing decision changes from P_2 to P_1 as the risk-aversion factor α increases.

The perceived risk model can be thought of as a simple (dis)utility model. It is possible to model risk disutility in other ways. For example, Kalelkar and Brinks (1978) proposed an empirical disutility function that was constructed by using a series of questions posed to decision makers. Erkut and Ingolfsson (2000) proposed an exponential disutility function to model risk aversion.

3.3 Risk on a hazmat transportation route

Up to this point, we discussed hazmat transport risk in general. Now we discuss the modeling of risk along an edge, and then along a route, of a transport network. In other words, we now move from point risk (risk due to accident at a given point) to linear risk (risk along an edge and route). Consider a road network $G = (N, E)$ with node set N and edge set E. The nodes correspond to the origin, the destination, road intersections, and population centers and the edges correspond to road segments connecting two nodes. (We note that one does not have to model population centers as nodes if one uses a GIS as discussed earlier.) We first focus our discussion on road transportation, and then move to hazmat transportation on rail.

Note that in the context of hazmat routing it is desirable that each point on an edge has the same incident probability and level of consequence (e.g., population density). Therefore, a long stretch of a highway that goes through a series

of population centers and farmland should not be represented as a single edge, but as a series of edges. Thus, a network to be used for hazmat routing is usually different from a network to be used for other transport planning purposes. This difference is quite important since it limits the portability of network databases between different transport applications (Erkut and Verter, 1998). We first discuss the modeling of risk along an edge.

3.3.1 Edge risk

Erkut and Verter (1998) proposed a risk model that takes into account the dependency to the impedances of preceding road segments (see also Jin et al., 1996; Jin and Batta, 1997). Suppose that an edge is a collection of n unit road segments each with the same incident probability p and consequence c. The probability p is obtained from (3.1) and the consequence c is determined by taking a proper impact area of a unit road segment. If, for example, the impact area of a unit road segment is modeled as a danger circle, then the impact area of an edge is a semicircular shape with the same radius as the danger circle, as shown in Figure 12. The vehicle will either have an incident on the first road segment, or it will make it safely to the second segment. If it makes it safely to the second segment, it will either have an incident in the second segment, or it will not, and so on. They assumed that the trip ends if an incident occurs. Hence, the expected risk associated with this edge would be

$$pc + (1-p)pc + (1-p)^2 pc + \cdots + (1-p)^{n-1} pc. \tag{3.6}$$

Since the incident probability p is at most on the order of 10^{-6} per trip per kilometer (based on North American data, Harwood et al., 1993), we can approximate

$$p^s \cong 0, \quad \text{for } s > 1. \tag{3.7}$$

Consequently, the risk of hazmat transport on this edge becomes pnc. For edge i, we can, thus, define the risk as

$$r_i = p_i c_i, \tag{3.8}$$

where the probability of an incident on edge i is $p_i := np$, and the associated consequence is $c_i := c$.

Note that this simple risk model works under an assumption of uniform incident probability and uniform consequence along an edge. If these two

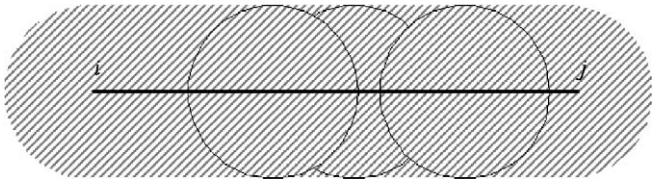


Fig. 12. Semicircular impact area around link (i, j).

attributes are not uniform, the risk computation on either an edge or an origin–destination route will be more complicated. In practice, however, this assumption will be valid if we define long stretches of a highway as a series of edges. (In other words, it is not only the network topology, but also the value of the edge attributes that define an edge. The edges must be short enough that the accident probability and the consequence are constant along the entire edge.) This edge risk definition can be considered as a generalization of the classical (or traditional) risk definition, which considers risk as an expected loss (see Section 3.1). The expected loss can be obtained from (3.6) by defining $n = 1$, i.e., each unit road segment is considered as an edge of the road network. Next we will discuss in detail some ways to model and calculate the edge risk.

Recall that according to Equation (3.2), the risk of a hazmat accident on road segment l can be calculated by considering the probability that individuals in neighborhood L (of road segment l) will be affected due to the incident and the population density in L. A hazmat vehicle at point v on edge (i, j) poses a threat to the population at each point v' in the impact area L. The hazmat incident probability $p_{ij}(v)$, can be obtained from (3.1) and it is measured in probability of accident per-unit length of movement. Moreover, let us assume that the consequence is determined by assuming that the impact area is a danger circle with radius λ.

The edge-risk formulation can be derived as follows. Let l_{ij} denote the length of edge (i, j) and $w_{v'}$ denote the population density at a point v'. The risk at point v', $r_{v',ij}$, due to the hazmat transport on an edge (i, j) is determined by

$$r_{v',ij} := w_{v'} \int_{v=0}^{l_{ij}} \delta(v, v')\, p_{ij}(v)\, \mathrm{d}v, \tag{3.9}$$

where

$$\delta(v, v') := \begin{cases} 1, & d(v, v') \leqslant \lambda, \\ 0, & \text{otherwise}, \end{cases}$$

with $d(v, v')$ the Euclidean distance of two points v and v'. To calculate the integral $\int_{v=0}^{l_{ij}} \delta(v, v')\, p_{ij}(v)\, \mathrm{d}v$, one can move the origin to node i and rotate the axes so that edge (i, j) lies on the positive abscissa. Denote this integral by $F_{ij}(vv')$. The semicircular area around an edge (i, j) consists of four regions with different expressions to calculate $F_{ij}(v')$, as shown in Figure 13. We note that region II is empty when $l_{ij} > 2\lambda$. If the coordinate of v' is $(x_{v'}, y_{v'})$ and $x^+(v')$ and $x^-(v')$ are the intersections of the abscissa with a circle of radius λ centered at v', then

$$x^+(v') = x_{v'} + \sqrt{\lambda^2 - y_{v'}^2} \quad \text{and} \quad x^-(v') = x_{v'} - \sqrt{\lambda^2 - y_{v'}^2}, \quad \text{if } \lambda > |y_{v'}| \tag{3.10}$$

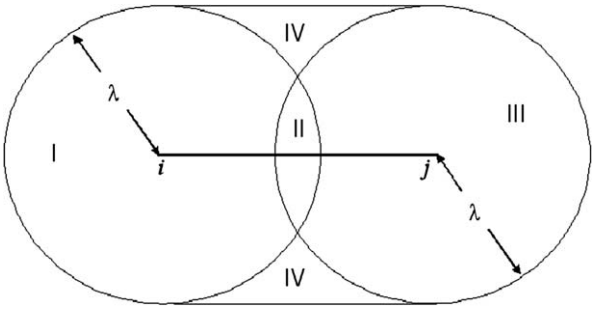


Fig. 13. Regions inside the semicircular impact area with radius λ around edge (i, j) (Batta and Chiu, 1988).

for every point v' in the road network. In this case, Batta and Chiu (1988) showed that

$$F_{ij}(v') = \begin{cases} \int_0^{x^+(v')} p_{ij}(v)\,\mathrm{d}v, & v' \text{ is in region I,} \\ \int_0^{l_{ij}} p_{ij}(v)\,\mathrm{d}v, & v' \text{ is in region II,} \\ \int_{x^-(v')}^{l_{ij}} p_{ij}(v)\,\mathrm{d}v, & v' \text{ is in region III,} \\ \int_{x^-(v')}^{x^+(vv')} p_{ij}(v)\,\mathrm{d}v, & v' \text{ is in region IV,} \\ 0, & v' \text{ is outside the semicircular area.} \end{cases} \tag{3.11}$$

Hence, the total risk of a hazmat vehicle travels on edge (i, j) is

$$r_{ij} = \int_{v' \in L} r_{v',ij}\,\mathrm{d}v'.$$

Batta and Chiu (1988) assumed that population centers are located at nodes and along the edges of the road network. Thus, a hazmat vehicle at point v on edge (i, j) poses a threat to the population at node v' and/or at point v'' on edge (i', j'). Let $w_{v'}$ denote the population density at node v', and $f_{kl}(v'')$ denote the population density function associated with edge (i', j'). Moreover, assume that the function $f_{i'j'}(v'')$ has been normalized so that its integral from zero to $l_{i'j'}$ equals one. The nodal risk at node v', $r_{v',ij}$, is determined by (3.9) and the edge risk on edge (i', j'), $r_{i'j',ij}$, due to the hazmat transport on edge (i, j) is determined by

$$r_{i'j',ij} := \int_{v''=0}^{l_{i'j'}} f_{i'j'}(v'') \int_{v=0}^{l_{ij}} \delta(v, v'')\, p_{ij}(v)\,\mathrm{d}v\,\mathrm{d}v''.$$

To calculate the edge risk $r_{i'j',ij}$, we need to partition edge (i', j') into regions as discussed earlier (see Figure 14). Let consider a point v'' on (i', j'), which is v'' units from node i'. By definition, the coordinates of this point are

$$x_{v''} = x_{i'} + (x_{j'} - x_{i'})\frac{v''}{l_{i'j'}} \quad \text{and} \quad y_{v''} = y_{i'} + (y_{j'} - y_{i'})\frac{v''}{l_{i'j'}}.$$

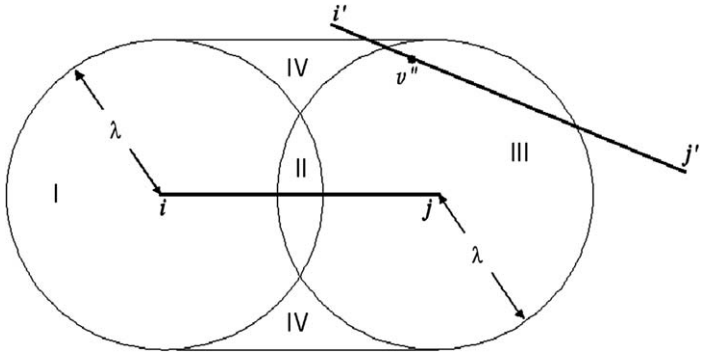


Fig. 14. Partition of edge (i', j') into regions inside the semicircular impact area with radius λ around edge (i, j).

Using this coordinate and (3.10) and (3.11), one can calculate $x^+(v'')$, $x^-(v'')$, $F_{ij}(v'')$, and finally the edge risk $r_{i'j',ij}$. Hence, the total risk of a hazmat vehicle travels on edge (i, j) is

$$r_{ij} = \sum_{(i',j')} r_{i'j',ij} + \sum_{v'} r_{v',ij}.$$

3.3.2 Path risk

An origin–destination route P for a hazmat shipment is a collection of edges, where travel on this path can be viewed as a probabilistic experiment as shown in Figure 15 (see, e.g., Jin et al., 1996; Jin and Batta, 1997; Erkut and Verter, 1998). Similar to the argument used above, a hazmat vehicle will travel along the ith edge of P only if there is no incident on the first $(i-1)$ edges of P (i.e., an incident terminates a trip along P). Suppose that the path P has n edges. Note that n may represents the length of the path if each edge $e_i \in E$, $i = 1, \ldots, n$, has length of one unit. The expected path risk associated with this trip can be expressed as

$$R(P) = \sum_{i=1}^{n} \prod_{j=1}^{i-1} (1 - p_j) p_i c_i, \tag{3.12}$$

where $(1-p_1)(1-p_2)\cdots(1-p_{i-1})p_i c_i$ is the impedance of the ith edge of P. By this definition, edge impedances are path-dependent.

Using an approximation similar to (3.7), that is, assuming

$$p_i p_j \cong 0, \quad \text{for all edges } i, j, \tag{3.13}$$

we obtain a very simple linear form of path risk

$$R'(P) = \sum_{i=1}^{n} p_i c_i. \tag{3.14}$$

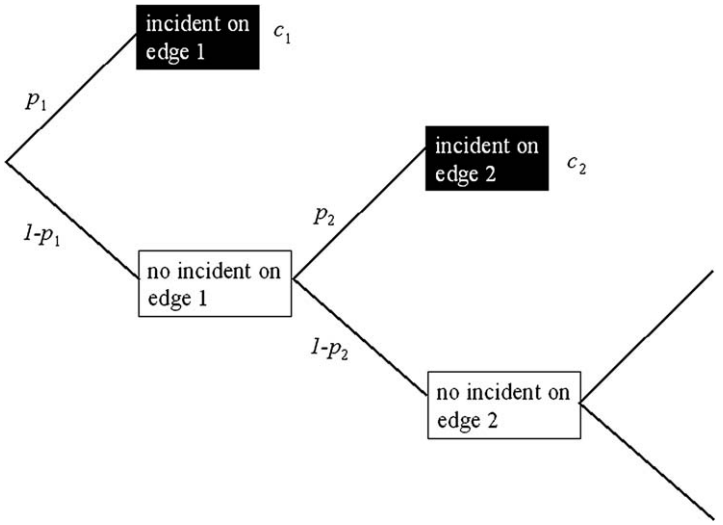


Fig. 15. Partial probability path displaying possible outcomes of a hazmat transport along a path, where p_i is the incident probability and c_i is the associated consequence along the ith edge of the path (Erkut and Verter, 1998).

This model is often referred to as the traditional risk model, since it explicitly uses the expected consequence definition of risk. Note that this model is simple to explain and justify, and it is not data intensive; it requires only one accident probability and one consequence figure per edge. Furthermore, it is rather easy to work with in optimization models. In fact, minimizing (3.14) for a given OD pair in a hazmat transport network is a shortest path problem which is solved easily for even large networks. For these reasons, most papers on hazmat transportation use this traditional risk model (Erkut and Verter, 1998). The US DOT also uses this approach in their guidelines (US DOT, 1994).

This simple risk model makes a tacit assumption that the hazmat vehicle will travel along every edge on the path, regardless of what happened on earlier edges. Consequently, a single hazmat trip can result in several incidents (with a very small probability). In some cases, though very unlikely, this assumption is practically reasonable. After a minor incident, the cargo may still be transported to the destination, perhaps on a different vehicle and/or on different route. Nevertheless, since incident rates for hazardous materials are very low, the probability of the conditioning event that an incident has not yet occurred when an edge i is reached will always be very close to 1. Therefore, the two assumptions (an incident terminates a trip and an incident does not terminate a trip) and (3.12) and (3.14), consequently, will differ insignificantly. Erkut and Verter (1998) point out that this approximation is likely to result in a very small error (less than 0.25% in most cases) in measuring the incidence probability along a hazmat transport route. Erkut and Ingolfsson (2000) provide an upper bound of $\exp(np_{\max}) - 1$ on the percent of error introduced by (3.14)

relative to (3.12), where $p_{\max}$ is an upper bound on the incident probability on any edge. Formally

$$\frac{R'(P) - R(P)}{R(P)} < \exp(np_{\max}) - 1.$$

For example, for a path P with length 4800 km, using an incident probability of 10^{-6} per trip-km, we can compute an upper bound of $\exp(np_{\max}) - 1 = \exp(0.004800) - 1$, which is 0.48%. This upper bound is obtained by assuming that accidents along any edge i with length l_i occur according to a spatial Poisson process with rate λ_i per distance unit. Under this assumption, the risk on path P can be obtained as

$$R''(P) = \sum_{i=1}^{n} \prod_{j=1}^{i-1} \exp(-p_j)\big(1 - \exp(-p_i)\big)c_i, \tag{3.15}$$

where $p_i = \lambda_i l_i$. By defining

$$q_i := 1 - \exp(-p_i) \tag{3.16}$$

for all edges i, (3.15) reduces to (3.12) with q_i replacing p_i.

Although the traditional risk model has been the most popular one, many other hazmat transport risk models have been proposed in the literature. Table 3 summarizes nine models and cites studies that have used each model. Each of the seven models that use probabilities are based on approximation (3.13), even though this approximation is usually not mentioned explicitly. We will refer to the alternate expressions of these seven models without using approximation (3.13) as "exact." Most of the models use population exposure as the consequence measure. In the population exposure model, c_i denotes the total population in the rectangle shape impact area that stretches along edge i. Other models use the circle-shaped impact area. Based on the empirical analysis on the US road network, Erkut and Verter (1998) suggest that the choice of risk model is important because it effects the path selection decision and the optimal path for a certain criterion can perform very poorly under another. Therefore, researchers as well as practitioners must pay considerable attention to the risk modeling in hazmat transport.

In addition to the path risk models summarized in Table 3, Jin and Batta (1997) proposed six exact risk models, which relate the number of shipments or trips S that need to be made and the threshold number of accidents T. The shipments cease after T accidents occur or S trips have been made, whichever come first. The hazmat shipments are considered as a sequence of independent Bernoulli trials. Moreover, it is assumed that a trip is over if an accident occurs on that trip or the destination is reached. Here, we provide a summary of these risk models and refer the reader to Jin and Batta (1997) for more detail.

- Expected consequence of each trip given that shipment will continue no matter how many accidents occur (i.e., when $S = T = \infty$).

Table 3.
Alternative models of path risk (adapted from Erkut and Ingolfsson, 2005)

Model	Approximation approach	Satisfy axioms? (approximation/ exact model) Y = yes; N = no; NA = not applicable			Sample references
		Axiom 1	Axiom 2	Axiom 3	
Traditional risk	$\sum_{i=1}^{n} p_i c_i$	Y/N	Y/N	Y/N	Batta and Chiu, 1988; US DOT, 1994; Alp, 1995; Zhang et al., 2000
Population exposure	$\sum_{i=1}^{n} c_i$	NA/Y	NA/Y	NA/Y	Batta and Chiu, 1988; ReVelle et al., 1991
Incident probability	$\sum_{i=1}^{n} p_i$	Y/Y	Y/Y	Y/Y	Saccomanno and Chan, 1985; Abkowitz et al., 1992
Perceived risk	$\sum_{i=1}^{n} p_i c_i^{\alpha}, \alpha > 0$	Y/N	Y/N	Y/N	Abkowitz et al., 1992
Conditional risk	$\sum_{i=1}^{n} p_i c_i / \sum_{i=1}^{n} p_i$	N/N	N/N	N/N	Sivakumar et al. 1993, 1995; Sherali et al., 1997
Maximum population exposure	$\max_{e_i \in P} c_i$	NA/Y	NA/Y	NA/Y	Erkut and Ingolfsson, 2000
Expected disutility	$\sum_{i=1}^{n} p_i(\exp(\alpha c_i) - 1), \alpha > 0$	Y/N	Y/N	Y/N	Erkut and Ingolfsson, 2000
Mean–variance	$\sum_{i=1}^{n} (p_i c_i + \beta p_i c_i^2)4, \beta > 0$	Y/N	Y/N	Y/N	Sivakumar and Batta, 1994; Erkut and Ingolfsson, 2000
Demand satisfaction	$\sum_{i=1}^{n}(1 - \exp(-p_i))c_i \prod_{j=i}^{n} \exp(p_j)$	NA/Y	NA/N	NA/Y	Erkut and Ingolfsson, 2005

Note: The three axioms tabulated here are discussed in the next subsection.

- Expected total consequence given that shipments will cease either when T accidents occur or S shipments are finished (i.e., when $T < S < \infty$).
- Expected total consequence given that shipments will cease when T accidents have occurred (i.e., when $T < \infty$ and $S = \infty$).
- Expected total consequence given that shipments will cease when T accidents have occurred (i.e., when $T < \infty$ and $S = \infty$) and parameters change after an accident.
- Expected consequence per trip given that shipments will cease when T accidents have occurred (i.e., when $T < \infty$ and $S = \infty$).

- Expected number of trips between two successive accidents.

Most of the exact expressions in Jin and Batta (1997) are too complicated for optimization purposes, and hence only the associated approximate models are of interest for practical purposes. Yet, approximations for the fourth and fifth models above are still not available. Further research on situation-specific models, such as the six listed above, is warranted.

We now discuss briefly the last three rows in Table 3, which are the most recently proposed hazmat transport risk models.

Expected disutility model. The disutility model incorporates the risk aversion of the society toward hazmat incidents, especially the catastrophic incidents (incidents with very large consequences). Erkut and Ingolfsson (2000) assumed that hazmat incidents occur according to a spatial, nonhomogeneous Poisson process defined over the edges of the network. Let N_i and X_i denote the number of hazmat incidents that occur on the ith edge and the number of people affected by a hazmat incident on the ith edge, respectively, of path P, where N_i has a Poisson distribution with a parameter p_i, the incident probability on ith edge of path P. We can thus define $X_i = c_i N_i$, where c_i denotes the associate population exposure. The disutility function is defined as $u(X) := \exp(\alpha X)$, where the constant $\alpha > 0$ is a measure of catastrophe aversion. The higher the values of α, the more extreme the catastrophe aversion. By assuming that a single trip can result in several incidents, the expected disutility for a path P can be obtained as $E(u(X)) = \exp[\sum_{i=1}^{n} p_i(\exp(\alpha c_i) - 1)]$. Minimizing $E(u(X))$ is then equivalent to minimizing the summation in the exponent, i.e., $\sum_{i=1}^{n} p_i(\exp(\alpha c_i) - 1)$. Hence, finding a minimum disutility path is equivalent to finding a shortest path with edge attribute $p_i(\exp(\alpha c_i)-1)$. The magnitude of the edge attributes can become very large. For example, suppose the population exposure is 10,000, the incident probability is 10^{-6}, and the risk aversion constant is 0.01. Then, the edge attribute is $10^{-6}(\exp(100)-1) \approx 10^{36}$. As the risk aversion constant α increases, the edge attribute will approach infinity. Consequently, this will ban the associated edge from consideration during a route selection process that seeks a finite solution. Under an assumption that an incident terminates the trip, the expected utility for a path P (i.e., the exact model) can be obtained as

$$E\big(u(X)\big) = \exp\left[\sum_{i=1}^{n} r_i\big(\exp(\alpha c_i) - 1\big)\right],$$

where

$$r_i := \prod_{j=1}^{i-1} \exp(-p_j)\big[1 - \exp(-p_i)\big], \quad e_i \in P, \tag{3.17}$$

denotes the incident probability on edge i conditioned that no incident occurred on the first $(i-1)$ edges. By definition, r_i are path dependent.

Mean–variance model. Many available risk models are based solely on the expected value of the risk and ignore how risk may deviate from the mean value. Sivakumar and Batta (1994) proposed a risk model that identifies the least expected length path subject to the constraint that the variance of the path length is within a pre-specified threshold. The model is formalized as an integer programming problem with linear objective function and both linear and nonlinear constraints. The nonlinear constraints contain quadratic terms which account for the covariance of length between two edges. Since the covariance terms can be negative, subtour elimination constraints are added to ensure a simple-path solution. The authors developed an efficient solution procedure, based on the Lagrange multipliers, to solve the equivalent linear integer programming problem, which is obtained by linearizing the quadratic terms.

Under the same Poisson distribution for the incident rates as in the disutility model, Erkut and Ingolfsson (2000) proposed a risk model that takes into account both the expected value and variance of the number of people affected by an incident. Using the same definition of X_i, N_i, and c_i, and assuming that a single trip can result in several incidents (i.e., the approximate model), the expected value and the variance of $X(P)$, the total number of people affected by incidents caused by travel along P, are $E[X(P)] = \sum_{i=1}^{n} c_i p_i$ and $\text{Var}[X(P)] = \sum_{i=1}^{n} c_i^2 p_i$. The associate exact models are $E[X(P)] = \sum_{i=1}^{n} c_i r_i$ and $\text{Var}[X(P)] = \sum_{i=1}^{n} c_i^2 r_i - (\sum_{i=1}^{n} c_i r_i)^2$, where r_i are defined as in (3.17). One can consider these two measures $E[X(P)]$ and $\text{Var}[X(P)]$ simultaneously in a multiobjective model and search for paths that are Pareto-optimal with respect to both $E[X(P)]$ and $\text{Var}[X(P)]$. To deal with the multiobjective model, one can use the weighted sum technique and obtain a disutility model $E[X(P)] + \beta \,\text{Var}[X(P)]$ for a given constant β. By minimizing this for several values of β, several Pareto-optimal paths can be found.

Demand satisfaction model. When a hazmat is transported to satisfy a demand (e.g., a shipment of chlorine from a producer to a chemical processing plant), an incident will result in a subsequent shipment. Hence, we must consider the possibility of multiple trips to fulfill the demand. Erkut and Ingolfsson (2005) proposed a simple demand satisfaction model by assuming that an incident will terminate a trip (i.e., referring to exact model) and a new shipment must be arranged to fulfill the demand. The exact probability that transport along a path P results in at least one incident is

$$\bar{p}(P) = 1 - \prod_{i=1}^{n} (1 - q_i),$$

where q_i are defined as in (3.16). By assuming that this probability is independent of any previous trips that were terminated by an incident, then one can consider each trip as a Bernoulli trial, with probability $1 - \bar{p}(P) = \prod_{i=1}^{n} (1 - q_i)$ of success in any given trial. The number of trips required (on the same path)

before the first success (i.e., trip arrives at the destination safely) will then follows a geometric distribution with expected value $1/\prod_{i=1}^{n}(1-q_i)$. By taking the expected consequence per trip as in (3.15), the expected total consequence from all trips required to fulfill demand is

$$R'''(P) = \frac{\sum_{i=1}^{n}\prod_{j=1}^{i-1}(1-q_j)q_ic_i}{\prod_{j=1}^{n}(1-q_j)} = \sum_{i=1}^{n} q_ic_i\prod_{j=i}^{n}(1-q_j)^{-1}. \qquad (3.18)$$

The expression in (3.18) has the following intuitive interpretation: the term q_ic_i is the expected risk associated with traversing edge i once and the term $\prod_{j=i}^{n}(1-q_j)^{-1}$ is the expected number of times that edge i and the subsequent edges on the path must be traversed before the shipment reaches the destination.

3.3.3 Path risk axioms

Now, we will discuss three important axioms which can be used to assess the merits of the different models listed in Table 3. Define $v(P)$ to be an evaluation function that operates on path P (such as distance, cost, or risk). Let $\mathbf{P}_1$ denote the set of all paths between an origin O_1 and a destination D_1, and $\mathbf{P}_2$ denote the set of all paths between an origin O_2 and a destination D_2. Let assume that for any $P_1 \in \mathbf{P}_1$ there is $P_2 \in \mathbf{P}_2$ such that $P_1 \subset P_2$.

Axiom 1 (Monotonicity axiom for path evaluation models (Erkut, 1995)). If a path P_1 is contained in a path P_2, then $v(P_1) \leqslant v(P_2)$.

Axiom 2 (Optimality principle for path selection models (Erkut and Verter, 1998)).

$$v(P_2) = \min_{P\in\mathbf{P}_2} v(P) \quad \Longrightarrow \quad v(P_1) = \min_{P\in\mathbf{P}_1} v(P).$$

For the third axiom, we assume that $v(P)$ is a function of K edge vector attributes $u_k(P)$ of size n, the number of edges in P, i.e., $v(P) = f(u_1(P), \ldots, u_K(P))$. For example, the attributes of any edge in P can be the incident probability and its associated consequence. In this case, we have $K = 2$.

Axiom 3 (Attribute monotonicity axiom (Erkut and Verter, 1998)). If h_k, $k = 1, \ldots, K$, are nonnegative vector of reals of size n, then

$$f\big(u_1(P), \ldots, u_K(P)\big) \leqslant f\big(u_1(P) + h_1, \ldots, u_K(P) + h_K\big).$$

The first axiom implies that the evaluation value of a path will not decrease as edges are added to the path. Clearly additive value functions (e.g., distance, cost, travel time) satisfy this monotonicity axiom. The second axiom is merely a restatement of Bellman's optimality principle that implies a concatenating

property of the shortest path. That is, all subpaths of an optimal path should themselves be optimal. Evaluation functions that satisfy Axiom 2 are called *order-preserving functions*. The third axiom states that the path evaluation function is a nondecreasing function of edge attributes. Consequently, path risk is a nondecreasing function of edge incident probabilities and edge consequences, i.e., increased probability or consequence on an edge cannot result in reduced path risk.

One of the nine models in Table 3, namely the conditional risk model, violates all three axioms. Erkut (1995) and Erkut and Verter (1998) argued that this model has some undesirable properties which make the model inappropriate for planning of hazmat shipments. For example, increasing the accident probability on a link may reduce the conditional risk of a route that includes that link.

We now consider the remaining eight models in Table 3. Most of the approximate versions of the models listed in Table 3 satisfy all three of these axioms. However, without assumption (3.7) or (3.13), most of the "exact" models containing probabilities do not satisfy the axioms. For example, consider the exact version of the traditional risk model defined in (3.12). One can easily construct a simple example to demonstrate that looping reduces the risk (see, e.g., Boffey and Karkazis, 1995; Erkut and Verter, 1998; Erkut and Ingolfsson, 2005). A loop in hazmat route is clearly undesirable (and indefensible). Therefore when using this exact model one must restrict the feasible set to loopless paths (as in Boffey and Karkazis, 1995). However, if one makes assumption (3.7), looping will not occur. Hence, the approximate version of the traditional risk model does not have the undesirable property of the exact version.

The simple example in Figure 16 demonstrates how the exact traditional risk model may result in an indefensible route selection. Node 1 is the origin and node 4 is the destination. The incident probabilities and consequences are given along the edges.

The exact risks associated with the two paths are as follows:

$$\begin{aligned}
&\text{Path}(1,2,4):\\
&\qquad 10^{-4}\times 10+\left(1-10^{-4}\right)\times 10^{-4}\times 110{,}000=10.9999,\\
&\text{Path}(1,3,4):\quad 1\times 10+0=10.
\end{aligned}$$

Hence, the exact version of the traditional risk model would select Path$(1,3,4)$, and this selection is guaranteed to result in an incident. The downstream consequences on edges $(2,4)$ and $(3,4)$ are so high that the model chooses the path which guarantees the truck will not reach the downstream edges. Such a model is not suitable for decision making.

In general, in spite of their more realistic assumption (i.e., an incident will terminate the trip) most of the exact versions of risk models have some puzzling properties and they may be unsuitable for hazmat transportation planning. We suggest that researchers and practitioners consider the properties of the risk models carefully before selecting one.

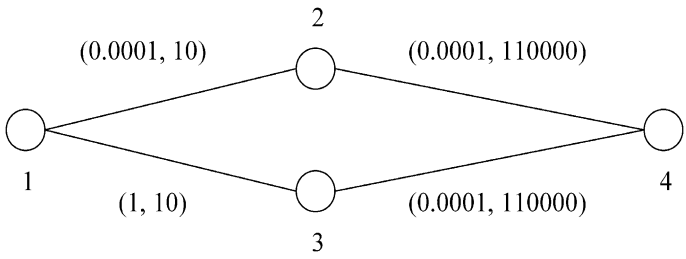


Fig. 16. Numerical example to demonstrate an undesirable property of the exact traditional risk model.

4 Routing and scheduling

Routing hazmat shipments involves a selection among the alternative paths between origin–destination pairs. From a carrier's perspective, shipment contracts can be considered independently and a routing decision needs to be made for each shipment, which we call the *local route planning* problem. A shipment typically involves multiple vehicles that have to be scheduled. Since the risk factors pertaining to each alternative route (such as accident probability and population exposure) can vary with time, the vehicle routing and scheduling decisions are intertwined, which we call the *local routing and scheduling problem*. At the macro level, hazmat routing is a "many to many" routing problem with multiple origins and an even greater number of destinations (List and Abkowitz, 1986). In the sequel, we refer to this problem as *global route planning*.

The local routing problem is to select the route(s) between a given origin–destination pair for a given hazmat, transport mode, and vehicle type. Thus, for each shipment order, this problem focuses on a single commodity and a single origin–destination route plan. Since these plans are often made without taking into consideration the big picture, certain links of the transport network tend to be overloaded with hazmat traffic. This could result in a considerable increase of accident probabilities on some road links as well as leading to inequity in the spatial distribution of risk. Although large-scale hazmat carriers are known to consider transport risk in their routing and scheduling decisions (Verter and Erkut, 1997), transport costs remain as the carriers' main focus.

In contrast, the government (municipal, state/provincial, or federal) has to consider the global problem by taking into account all shipments in its jurisdiction. This leads to a harder class of problems that involve multicommodity and multiple origin–destination routing decisions. In addition to the total risk imposed on the public and environment, a government agency may need to consider promoting equity in the spatial distribution of risk. This becomes crucial in the event that certain population zones are exposed to intolerable levels of risk as a result of the carriers' routing and scheduling decisions. The governments' task is further complicated by the need to keep the transport sector

economically viable – despite the regulations to ensure public safety – since dangerous goods shipments are an integral part of our industrial lifestyle.

Hazmat local route planning has attracted the attention of many OR researchers. The existing local route planning models cover a wide area that includes different transport modes: road (e.g., Akgün et al., 2000; Kara et al., 2003), rail (e.g., Glickman, 1983; Verma and Verter, 2007), water (e.g., Iakovou et al., 1999; Iakovou, 2001); deterministic (e.g., Batta and Chiu, 1988; ReVelle et al., 1991) or stochastic models (e.g., Miller-Hooks and Mahmassani, 1998; Erkut and Ingolfsson, 2000); and single objective (Erkut and Verter, 1998; Erkut and Ingolfsson, 2005) or multiple objective models (e.g., Sherali et al., 1997; Marianov and ReVelle, 1998). Tables 2(a–d) provides a more complete list of references.

The local routing models fail to capture the dynamic nature of transport risk factors at the tactical level (e.g., traffic conditions, population density, and weather conditions). Moreover, most of these risk factors cannot be known a priori with certainty. They are both time-dependent and stochastic in nature; i.e., they are random variables with probability distribution functions that vary with time. Therefore, the local routing and scheduling problem is best modeled as a path selection problem in a *stochastic time-varying* network (see, for example, Hall, 1986; Fu and Rilett, 1998; Miller-Hooks and Mahmassani, 1998; Miller-Hooks, 2001).

The global route planning problem has attained relatively little attention in the literature, much less compared to the local route planning problem. The results in this area include the works of Gopalan et al. (1990b), Lindner-Dutton et al. (1991), Marianov and ReVelle (1998), and Iakovou et al. (1999). The works of Akgün et al. (2000) and Dell'Olmo et al. (2005) on the problem of finding a number of spatially dissimilar paths between an origin and a destination can also be considered in this area.

The rest of this section provides a discussion on the known models and solution algorithms pertaining to the three problem categories discussed above.

4.1 Local routing problems

As we have discussed in Section 3, almost all approximate versions of the path evaluation functions listed in Table 3 are additive and satisfy the optimality principle (i.e., Axiom 2). Therefore, the static, deterministic and single objective local routing problems that minimize those evaluation functions reduce to the classical shortest path problem. Consequently, a label-setting algorithm (e.g., Djikstra's algorithm) can simply be applied to find an optimal route.

Most of the exact versions of these path evaluation functions, on the other hand, do not satisfy Axiom 2. Therefore, Djikstra's algorithm cannot be applied directly to find an optimal route. Kara et al. (2003) proposed a simple modification of Djikstra's algorithm to find a route that minimizes the exact version of the path incident probability. The modification relies on the adjustment of the link attribute that is used to update the node label and

the scanning process. The algorithm is called the *impedance-adjusting node labeling shortest path algorithm* and is explained briefly as follows. Let $P = \{(i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$ with i_1 the origin node and i_n the destination node, and let $q(i_k)$ denote the probability of safely arriving at node i_k of P. From (3.12), we obtain $q(i_k) = q(i_{k-1})(1 - p_{i_{k-1}i_k})$ for $k = 2, \ldots, n$, where $p_{i_{k-1}i_k}$ denotes the incident probability of (i_{k-1}, i_k) and $q(i_1) := 1$. The attribute a_{ij} of each link (i, j) is defined as $a_{ij} := q(i)p_{ij}$. During the scanning process of node i, a_{ij} for each (i, j) is recomputed. This new value is used to update the node label θ_j of node j. If the current value of θ_j is greater than $\theta_i + a_{ij}$, then we set $\theta_j := \theta_i + a_{ij}$, $q(j) := q(i)(1 - p_{ij})$ and update the predecessor of node j. This modified algorithm has the same computational complexity as that of Djikstra's. Kara et al. (2003) also proposed the *impedance-adjusting link labeling* algorithm to minimize the path population exposure. This algorithm eliminates the errors resulting from double-counting of population exposure, which is caused by the network topology. Using a similar modification technique to the *impedance-adjusting node labeling shortest path algorithm*, Djikstra's algorithm can be used to solve the local routing problem with the exact version of perceived risk, the expected disutility, and the mean-variance path evaluation functions.

4.1.1 Rail transport

A significant majority of the literature on hazmat routing focus on road shipments. This is not surprising, since trucks account for the largest percentage of hazmat shipments, as discussed in Section 1. Although train shipments can reach comparable levels to truck shipments from a total tonnage perspective (particularly in Europe and Canada), they received considerably less attention from researchers. Remarkably, the literature on marine, air, and pipeline transport of dangerous goods is in its infancy. McClure et al. (1988) pointed out a number of differences between rail and highway routing of hazmat transportation. Rail infrastructure is typically owned and maintained by private rail companies. Consequently, railroad networks are sparse and do not contain as many potential alternative routes as highway networks. More importantly, railroads do not have tracks circumventing major population centers that are comparable to interstate beltways around metropolitan areas. A given shipment is likely to be handled by more than one railroad carrier, whereas truck shipments are usually limited to a single company. The rail carriers are motivated to maximize their portion of the movement. In a recent paper, Verma and Verter (2007) highlighted additional differences between the two transport modes. A train usually carries nonhazardous and hazardous cargo together, whereas these two types of cargo are almost never mixed in a truck shipment. Furthermore, a rail tank car has roughly three times the capacity of a truck-tanker (80 tons and 25–30 tons respectively) and the number of hazmat railcars varies significantly among different trains. Another important characteristic of trains is the possibility of incidents that involve multiple railcars. Verma and Verter (2007) noted that there is an average of about one major railroad ac-

cident per year during the 1990–2003 period in the United States. Thus, there is a need for the development of risk assessment and routing procedures that incorporate the differentiating features of railroad hazmat shipments.

The academic literature has mostly focused on the comparison of rail and road from the viewpoint of hazmat transport risk. For example, Glickman (1988) observed that the accident rate for significant spills (when release quantities exceed 5 gallons or 40 pounds) is higher for truck tankers than for rail tank cars and that rail tank cars are more prone to small spills. Saccomanno et al. (1989) showed that the safer mode varies with the hazmat being shipped and differing volumes complicate comparison between the two transport modes. Leeming and Saccomanno (1994) reported that although hazmat railway shipments pose more risk to residents in the vicinity of railroad tracks, the total risk of these two transport modes does not differ significantly. Their conclusion is based on a single case study in England. In summary, there is no consensus among researchers with regards to the dominant transport mode in terms of public and environmental safety.

Over the past three decades, railroad industry has focused on reducing the frequency of tank car accidents as well as the likelihood of releases in the event of an accident – rather than routing and scheduling of trains with potentially hazardous cargo. The industry's most recent initiatives have aimed at improving tank car safety at the design stage. By studying the risks associated with nonpressurized materials, Raj and Pritchard (2000) report that the DOT-105 tank car design constitutes a safer option than DOT-111. Barkan et al. (2000) showed that tank cars equipped with surge pressure reduction devices experienced lower release rates than those without this technology. Barkan et al. (2003) undertook a study to identify proxy variables that can be used to predict circumstances most likely to lead to a hazmat release accident. They concluded that the speed of derailment and the number of derailed cars are highly correlated with hazmat release.

4.2 Multiobjective approaches to local routing

As discussed in Section 1, hazmat transportation is *multiobjective* in nature with multiple stakeholders. In general, there is no solution that simultaneously optimizes all the conflicting objective functions in a multiobjective problem. Instead, a set of *nondominated* solutions (or *Pareto-optimal* solutions) can be determined. A Pareto-optimal solution is one where we cannot improve on an objective without worsening at least one other objective. Local route planning often involves finding the set of Pareto-optimal routes between a given origin–destination pair. In the event that the decision maker's preferences among the conflicting objectives are available in advance, the problem can be reduced to a single objective optimization problem (via utility theory). The *most preferred solution can then be identified* among the Pareto-optimal solutions so as to maximize the preference function of the decision maker.

Nembhard and White (1997) considered the problem of determining the most preferred path that maximizes a multiattribute, nonorder-preserving value function both with and without intermediate stops. For the no-stop case, the problem is solved approximately by applying the dynamic programming algorithm as if a subpath of an optimal path were always optimal (i.e., by using an approximate method on the exact problem). The intermediate-stop case is solved approximately by approximating the nonorder preserving criterion with the linear order-preserving criterion and by properly applying the dynamic programming algorithm (i.e., by using an exact method on the approximated problem). Marianov and ReVelle (1998) proposed a linear optimization model to find the routes that minimize both the cost and the exact version of the probability of accident. The weighted sum technique is used to solve the biobjective problem and to approximate the set of Pareto-optimal routes. The associated weighted, single objective problem can thus be solved by simply applying the classical shortest path algorithm. Tayi et al. (1999) dealt with the cost equity and feasibility problem in hazmat routing, where each edge of the network is associated with a vector of costs incurred by different zones due to an accident along that edge. The zones represent the community clusters, and each component of the cost vector represents the impact of an accident on a zone. The notion of cost equity is represented by six objective functions, including minimization of the average cost path, the maximum cost path, and the imbalanced cost path.

As discussed earlier, many transport risk factors involve considerable *uncertainty*, which increases the difficulty of routing decisions. Two methods that are frequently used in incorporating uncertainty are *mean-risk* (e.g., Markowitz, 1987; Ogryczak and Ruszczynski, 2002) and *stochastic dominance* (e.g., Yitzhaki, 1982; the survey by Levy, 1992). The mean-risk criterion is based on comparing only two values: the mean, representing the expected outcome; and the risk, a scalar measure of the variability of outcomes (e.g., variance and semivariance). Mean–variance (MV) criterion is probably the most well-known mean-risk criterion. It states that if $E(v(P_1)) \leqslant E(v(P_2))$ and $\mathrm{Var}(v(P_1)) \leqslant \mathrm{Var}(v(P_2))$ with at least one strict inequality, then $v(P_1)$ is MV-strictly smaller than $v(P_2)$, where $v(P)$ is an evaluation function that operates on path P.

Stochastic dominance (SD) criterion, on the other hand, considers the entire probability distribution rather than just the two moments. It uses the cumulative distribution function (CDF) as a basis for comparison. Let F_{P_1} and F_{P_2} be the CDFs of two random variables $v(P_1)$ and $v(P_2)$. The first- and second-order stochastic dominance (FSD and SSD) are defined as follows. A random variable $v(P_1)$ is strictly smaller, with respect to FSD, than a random variable $v(P_2)$, if $F_{P_1}(t) \geqslant F_{P_2}(t)$ for all values of t, and at least one of the inequalities holds strictly. If two CDFs do not intersect, then one of them should stochastically dominate the other, regardless of their variances. Furthermore, a random variable $v(P_1)$ is strictly smaller, with respect to SSD, than a random variable $v(P_2)$, if $\int_{-\infty}^{t} (F_{P_1}(\omega) - F_{P_2}(\omega))\,\mathrm{d}\omega \geqslant 0$ for all values of t, and at least one the

inequalities holds strictly. For SSD, the two CDFs may intersect, but the total accumulated area between F_{P_1} and F_{P_2} must stay nonnegative up to any t. FSD implies SSD but not vice versa.

Figure 17(a) shows that the distribution F_{P_1} is above distribution F_{P_2} everywhere, and therefore, the probability of "*t or less*" is higher under F_{P_1} than F_{P_2}. In Figure 17(b), if the two distributions cross within the range of t, then the FSD does not hold, but SSD holds. Figure 17(c) shows that $v(P_1)$ is neither FSD nor SSD smaller than $v(P_2)$ and vice versa. Mean–variance criterion offers a much simpler computational tool than SD criterion. However, a Pareto-optimal solution with respect to the MV criterion may be stochastically dominated by other feasible solutions if the normality of distributions is not guaranteed (see, e.g., Yitzhaki, 1982; Ogryczak and Ruszczynski, 2002). On the other hand, as the CDFs of $v(P_1)$ and $v(P_2)$ (or their integration) have to be compared for every t, the stochastic dominance itself is actually a multiobjective model with a continuum of criteria. The stochastic dominance criterion usually leads to large efficient sets, and it does not provide us with a simple computational tool.

The problem with the efficient set becomes worse in the multiobjective routing problem, as the number of Pareto-optimal solutions can be exponential in the number of nodes (Hansen, 1980). To reduce the size of this efficient set,

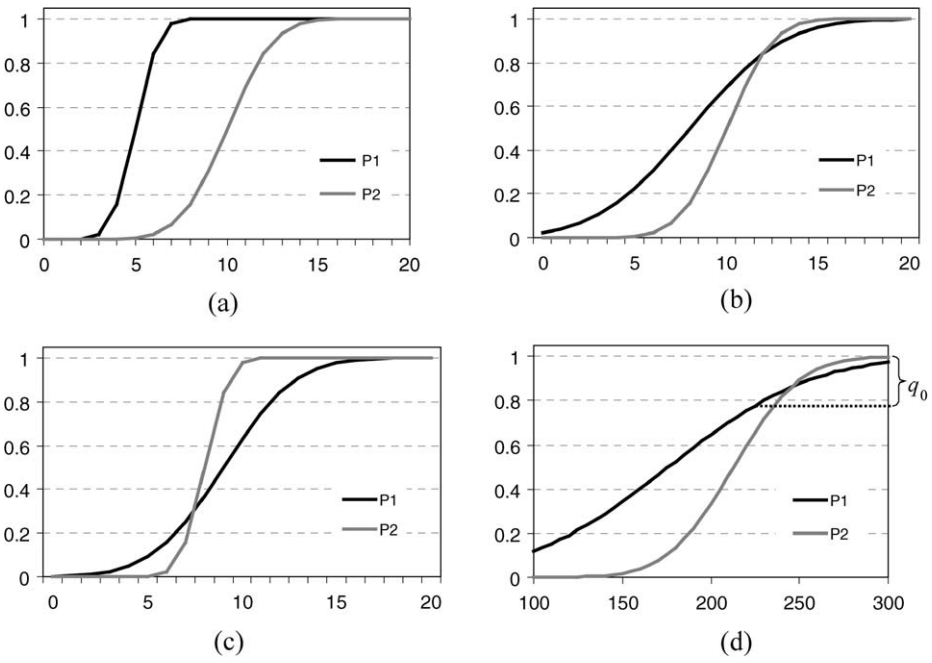


Fig. 17. (a) P_1 is FSD P_2; (b) P_1 is not FSD P_2 but P_1 is SSD P_2; (c) P_1 is not SSD P_2 and vice versa; (d) P_1 dominates P_2 for $q_0 \geqslant 0.1382$.

Wijeratne et al. (1993) proposed a two-stage evaluation procedure for normally distributed path evaluation functions. This procedure includes a probability parameter that allows the analyst or the decision maker to control the degree to which a comparison deviates from the FSD criterion. That is, a path P_1 dominates P_2 if either of the following occurs:

- *Primary comparison rule*: both the mean and the variance of $v(P_1)$ are smaller than those of $v(P_2)$ (i.e., MV criterion).
- *Secondary comparison rule*: the mean of $v(P_1)$ is smaller, the variance of $v(P_2)$ is smaller, and the CDF of $v(P_1)$ exceeds the CDF of $v(P_2)$ for probability values greater than $(1 - q_0)$.

The higher the value of q_0, the smaller the size of the efficient set. However, a small set may exclude some interesting Pareto-optimal paths. Consider the following small example. Suppose there are two paths P_1 and P_2 from an origin to a destination, where the mean and standard deviation of $v(P_1)$ and $v(P_2)$ are $(176; 64)$ and $(213; 30)$, respectively. Figure 17(d) shows that MV and FSD criteria do not hold. By applying the criterion of Wijeratne et al. (1993), path P_1 will dominate P_2 for $q_0 > 0.1382$.

Although the example involves a single evaluation function $v(P)$, observe that the incorporation of uncertainty results in a multiobjective problem. Wijeratne et al. (1993) proposed a simple procedure to deal with a stochastic multiobjective routing problem (or with a mixture of deterministic and stochastic path evaluation functions). We illustrate this procedure by a small example. Suppose that there are two stochastic path evaluation functions $v_1(P)$, $v_2(P)$, one deterministic path evaluation function $v_3(P)$ (all of these functions are to be minimized) and 4 paths to be compared. Hence, the set of feasible paths is $\mathcal{P} = \{P_1, P_2, P_3, P_4\}$. The comparison is done separately for each evaluation function, where the user-controlled probability parameter q_0 may be different for each stochastic evaluation function. Suppose we find (after applying the two-stage evaluation procedure) that with respect to $v_1(P)$, the set $\mathcal{P}$ can be partitioned into a ranked set $\mathcal{P}^1 = \{(P_1), (P_2, P_3), P_4\}$, which means P_1 dominates all other paths, P_2 is indifferent to P_3, and both P_2 and P_3 dominate P_4. Furthermore, with respect to $v_2(P)$, the set $\mathcal{P}$ may be partitioned into a ranked set $\mathcal{P}^2 = \{(P_1, P_3), P_4, P_2\}$. Suppose that $v_3(P_1) = 100$, $v_3(P_2) = 150$, $v_3(P_3) = 50$, and $v_3(P_4) = 100$, resulting in $\mathcal{P}^3 = \{P_3, (P_1, P_4), P_2\}$. We can thus combine the relative ranking for each path to create a ranking vector of evaluation functions for each path: Path $P_1 : (1, 1, 2)$; Path $P_2 : (2, 3, 3)$; Path $P_3 : (2, 1, 1)$; Path $P_4 : (3, 2, 2)$. (For the deterministic evaluation function, one may put its value, instead of the relative ranking directly in the ranking vector.) The final step is to examine this set of ranking vectors to eliminate the dominated paths. If we require strict dominance across all evaluation functions, we obtain two Pareto-optimal routes: P_1 and P_3.

Turnquist (1993) assumed that both accident probability and population exposure are stochastic. He studied the problem of identifying a set of Pareto-optimal routes with the following objectives: minimize the incident rate; min-

imize the population exposed within a certain distance of the roadway; and minimize the travel distance. Turnquist used the distribution functions of each Pareto-optimal path on each criterion to highlight the trade-offs among the Pareto-optimal solutions.

There are very few static and stochastic routing models (either single or multiobjective) in the literature for hazmat transportation. In addition to Wijeratne et al. (1993) and Turnquist (1993), the mean–variance models proposed by Sivakumar and Batta (1994) and Erkut and Ingolfsson (2000) are noteworthy (see the discussion on these papers in subsection "Mean–variance model" of Section 3.3.2). There are static, stochastic path finding models that are designed for other transportation applications (e.g., Frank, 1969; Mirchandani, 1976; Kulkarni, 1986; Corea and Kulkarni, 1993), which the reader may find useful. Nonetheless, the *dynamic*, stochastic routing is more relevant to hazmat transportation, which we discuss in the next section.

4.3 Local routing and scheduling problems

The traffic conditions and other risk factors in hazmat transportation networks (e.g., incident probabilities and population exposure) often vary with time and can at best be known a priori with uncertainty. For example, for a hazmat truck, the travel time and the accident probability on certain road segments can be uncertain and depend on traffic congestion, weather conditions, and road conditions during the vehicle's trip across those links. Hence, the transport risk and arrival time at the destination can vary with the dispatch schedule from the origin. Also, allowing the vehicle to stop during its trip in order to avoid peak risk periods on certain road segments can be an effective strategy to reduce the total transport risk (Erkut and Alp, 2006). To represent this phenomenon appropriately, the transport network should be modeled as a stochastic, time-varying (STV) network.

In an STV network, the link attributes (such as travel times, incident probabilities, and population exposure) are represented as random variables with a priori probability distributions that vary with time. STV network-based modeling has been an important and well-researched topic since the late 1980s (see, e.g., Hall, 1986; Fu and Rilett, 1998; Miller-Hooks and Mahmassani, 1998; Miller-Hooks, 2001). Most of the existing results are devoted to the Intelligent Transportation System (ITS), and only some of them are designed specifically for the hazmat transportation problem (e.g., Bowler and Mahmassani, 1998; Miller-Hooks and Mahmassani, 1998). The prevailing studies can be classified into three different groups:

1. *A priori optimization*: the optimal routes are chosen before the travel begins. Hence, an update on the routing decision en-route is not allowed.
2. *Adaptive route selection*: the routing decision is subject to change en-route based on the realization of the estimated data.

3. *Adaptive route selection with real-time updates*: the routing decision is subject to change en-route due to real-time updates of the traffic data followed by re-optimization procedures.

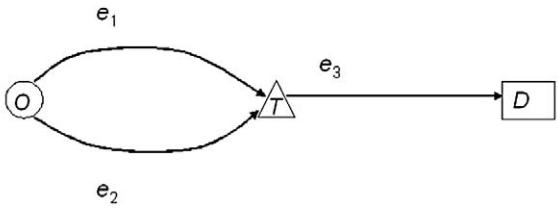
In the following, we will discuss some of the results in each class that can be applied to hazmat transportation.

4.3.1 A priori optimization

This class of problems assumes that the optimal route is chosen before trip begins. Hence, an update on the routing decision en-route is not allowed. All routing decisions in static (time-invariant) networks fall into this category.

Hall (1986) showed that in STV networks, one cannot simply set the random arc travel times to their expected values and identify the shortest (expected) travel time by applying standard shortest path algorithms based on Bellman's equation (Bellman, 1958), such as Djikstra's algorithm. The expected travel time on an arc in STV networks depends on the arrival time of the vehicle at the beginning of that arc. A partial route with a higher expected travel time might be selected, if this choice results in a preferable outcome in the rest of the route. This is demonstrated by the numerical example in Figure 18.

The objectives are to minimize the expected total travel time and to minimize the expected total risk as defined by the expected total number of exposed individuals. Suppose the hazmat truck must leave node O at 15:00. On the way to node T, arc e_1 has both, the lowest expected travel time, and the lowest expected population exposure (120 minutes and 100 individuals as opposed to 125 minutes and 120 individuals on arc e_2). Hence, Bellman's principle would include arc e_1 in the optimal path. However, note that a vehicle traveling on



Arcs	Travel times	Exposed populations
e_1	120 minutes (probability 1.0)	100
e_2	90 minutes (probability 0.3) 140 minutes (probability 0.7)	120
e_3	60 minutes (arrived at node *T* before 16:45) 120 minutes (otherwise)	50 (arrived at node *T* before 16:45) 200 (otherwise)

Fig. 18. An STV network for the fastest and least risk path problem.

arc e_2 has a higher probability of arriving at node T before 16:45 (0.3 probability as opposed to zero probability on arc e_1). Hence, the total expected travel time and the total expected population exposure via e_2 are lower ($0.3(90 + 60) + 0.7(140 + 120) = 227$ minutes and $0.3(120 + 50) + 0.7(120 + 200) = 275$ individuals as opposed to 240 minutes and 300 individuals).

Hall (1986) proposed an exact, nonpolynomial algorithm that combines a branch-and-bound technique with a k-shortest paths algorithm to find the fastest path in STV networks. This algorithm, however, applies only to acyclic networks or to cyclic networks with First-In First-Out (FIFO) travel times. (We say that travel times are FIFO if they are nondecreasing functions of time; i.e., if Vehicle A leaves before Vehicle B, Vehicle A will arrive no later than Vehicle B.) Miller-Hooks and Mahmassani (2000) extended Hall's model to allow cycles or non-FIFO travel times. They proposed a time-dependent label-correcting algorithm to solve this fastest path problem. Under the assumption that travel times are continuous functions of time, Fu and Rilett (1998) proposed a heuristic algorithm based on the k-shortest path algorithm to solve the fastest path problem without the FIFO assumption. The differentiating feature of their model is the propagation of mean and variance of travel time along a path in the process of determining the fastest path.

Chang et al. (2005) adapted the continuous-time mean and variance propagation method of Fu and Rilett (1998) to discrete-time intervals and minimized the total cost as well as the total travel time. The path evaluation functions (except the total travel time) of two paths in STV networks are comparable at a node only if the arrival times of those paths at this node are the same. This condition, however, implies a large efficient set, as it may be unlikely that two paths arrive at a node at precisely the same time. To tackle this problem, Sulijoadikusumo and Nozick (1998) and Chang et al. (2005) suggested a time-window criterion: two paths are comparable only if their arrival times are "close enough" as defined by the analyst/decision maker. Suppose Y_i^{Pj}, the arrival time at node i along a path P_j, is normally distributed. The probability that the difference of two path travel times is less than or equal to a predefined time window Δ can be approximated as

$$p\left(\left|Y_i^{P_1} - Y_i^{P_2}\right| \leqslant \Delta\right) = \Phi\left(\frac{\Delta - (E[Y_i^{P_1}] - E[Y_i^{P_2}])}{\sqrt{\mathrm{Var}[Y_i^{P_1}] + \mathrm{Var}[Y_i^{P_2}]}}\right) - \Phi\left(\frac{-\Delta - (E[Y_i^{P_1}] - E[Y_i^{P_2}])}{\sqrt{\mathrm{Var}[Y_i^{P_1}] + \mathrm{Var}[Y_i^{P_2}]}}\right),$$

where $\Phi(z)$ denotes the cumulative distribution function of a standard normal random variable. If $p(|Y_i^{P_1} - Y_i^{P_2}| \leqslant \Delta) \geqslant \delta$, where δ is the pre-specified threshold, then these two paths are comparable at node i. If the two paths are comparable, then the stochastic comparison methods discussed in the previous subsection can be used to choose the preferred path.

4.3.2 Adaptive route selection

When traveling along a network, the motorist gathers new information that can be useful in making better routing decisions. For example, the arrival time at a node can be used in making a choice among the partial emanating from that node. This is called *adaptive route selection*. The optimal route depends on intermediate information concerning past travel times, road and weather conditions and hence, a single (and simple) path is not adequate.

Hall (1986) showed that the optimal adaptive route in STV networks that minimizes the expected travel time is not a simple path but an acyclic subnetwork (called a *hyperpath*) that represents a set of routing strategies (see, e.g., Nguyen and Pallottino, 1986). The adaptive route specifies the road link to be chosen at each intermediate node, as a function of the arrival time at the node. As an illustration, consider the example depicted in Figure 19.

The hazmat truck is to leave node O at 15:00. The a priori expected travel times of two paths $P_1 := \{e_1, e_2\}$ and $P_2 := \{e_1, e_3\}$ are $0.3(90+60)+0.7(140+120) = 227$ minutes and $0.3(90 + 100) + 0.7(140 + 30) = 176$ minutes, respectively. The associated a priori expected total risks for P_1 and P_2 are $0.3(100+100)+0.7(150+80) = 221$ and $0.3(100+200)+0.7(150+50) = 230$ individuals at risk, respectively. Hence, the a priori fastest path is path P_2, and the a priori least risk path is P_1. However, if the motorist is permitted to select the rest of the path upon arrival at node T, we will obtain the following routing strategy:

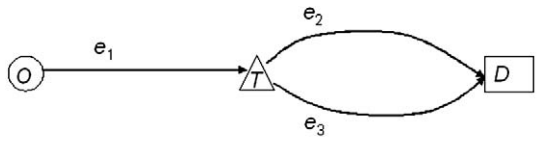
- Travel time: If the arrival time at node T is 16:30, take arc e_2 with a travel time of 60 minutes. If the arrival time at node T is 17:20, take arc e_3 with a travel time of 30 minutes. The expected travel time for the adaptive fastest path from O to D is $0.3(90+60)+0.7(140+30) = 164$ minutes. The associated total risk is $0.3(100 + 100) + 0.7(150 + 50) = 200$ individuals at risk.
- Total risk: The routing strategy is the same as for that of the adaptive fastest path.

The resulting hyperpath of the optimal adaptive routing strategy, depicted as a decision tree, is shown in Figure 20.

It is, in general, quite unlikely that the optimal adaptive routing strategies of different objectives coincide. In this case, the multiobjective version of the label correcting and Stochastic Decreasing Order of Time (SDOT) algorithms from Miller-Hooks (2001) can be used to generate a set of Pareto-optimal adaptive routing strategies.

4.3.3 Adaptive route selection with real-time updates

The recent advances in information and communication technologies, such as satellite-based Automatic Vehicle Location (AVL) and mobile phones, enable the driver and dispatch center to obtain and exchange real-time information. Satellite-based AVL is a computer-based vehicle tracking system that uses signals from satellite systems, such as Navstar Global Positioning System



Arcs	Travel times	Exposed populations
e_1	90 minutes (0.3) 140 minutes (0.7)	100 150
e_2	60 minutes (arrived at node *T* before 16:45) 120 minutes (otherwise)	100 (arrived at node *T* before 16:45) 80 (otherwise)
e_3	100 minutes (arrived at node *T* before 16:45) 30 minutes (otherwise)	200 (arrived at node *T* before 16:45) 50 (otherwise)

Fig. 19. An STV network for the adaptive routing strategy.

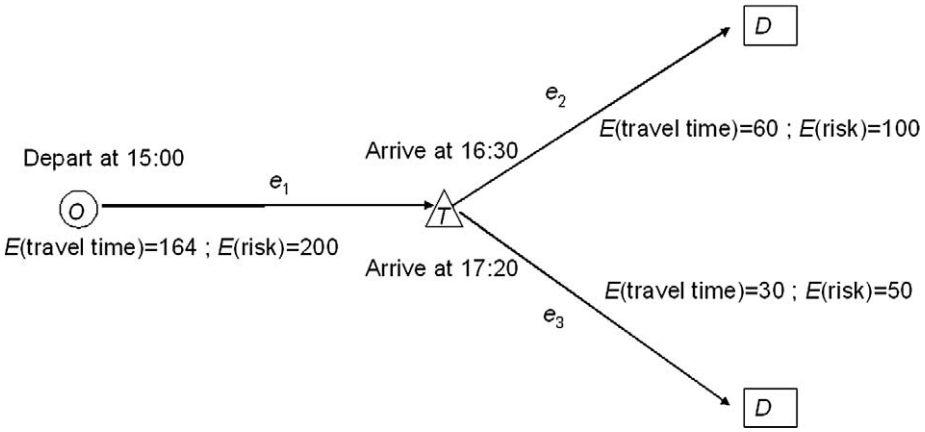


Fig. 20. The resulting hyperpath of the adaptive routing strategy, depicted as a decision tree.

(GPS), to identify a vehicle's location. Mobile communication systems such as cellular phones, paging systems, and mobile satellite communication systems, provide two-way communication between the driver and the dispatch center or among drivers. These AVL and mobile communication systems enable the driver and the dispatch center to monitor and/or change the route of vehicles based on real-time information.

These technological advances are challenging OR researchers to develop routing models and robust optimization procedures that are able to respond quickly to changes in the data. In this real-time environment, the quality of the decision depends not only on the appropriateness of the decision, but also on its timeliness (Seguin et al., 1997). Another main issue in this area, besides

route planning, is the data updating procedure. New real-time information obtained by the dispatch center is used to update the estimation of future values of some network attributes (e.g., travel times, incident probabilities, and population in the impact area). However, this information is of limited use if the information is about parts of the transport network that are far away from the current location of the vehicle (either spatially or temporally). Therefore, either a spatial or a temporal discounting procedure must be applied before this real-time information is used to update the estimates of network attributes (see, e.g., Hoffman and Janko, 1990; Koutsopoulos and Xu, 1993; Yang, 2001).

We observe a lack of papers in this area that consider both adaptive routing decisions and data updates based on real-time information. Moreover, none of the prevailing studies are designed specifically for hazmat transportation problems. Koutsopoulos and Xu (1993) proposed an information discounting procedure for travel times in finding the shortest path in STV networks with an FIFO assumption. For temporal discounting, they used the results from Hoffman and Janko (1990), where the ratio of the historical mean over its current travel time is used to estimate the future travel times on the same arc. Suppose that the route planning is defined in discrete time $\mathcal{T} := \{t_k : k = 0, \ldots, K\}$ with $t_{k+1} := t_k + \Delta$, $k = 0, \ldots, K-1$. If we denote the travel time ratio on arc (i, j) at time $t \in T$ by $\delta_{ij,t}$, then

$$\delta_{ij,t} := \frac{\bar{\lambda}_{ij,t}}{\lambda_{ij,t}},$$

where $\bar{\lambda}_{ij,t}$ is the historical average travel time on (i, j) at time t and $\lambda_{ij,t}$ is the associated actual travel time. This ratio is set to 1.0 when real-time information for an arc is not available. To incorporate changes in neighboring arcs, a smoothed mean ratio is computed as

$$\delta'_{ij,t} := \frac{1}{|A_{ij}|} \sum_{(k,l)\in A_{ij}} \delta_{kl,t},$$

where A_{ij} is a set of all adjacent arcs of (i, j). The new estimated travel time $\lambda'_{ij,t'}$ on arc (i, j) at a future time period $t' = t + \Delta t, \ldots, t_K$ is then given by

$$\lambda'_{ij,t'} := \frac{\bar{\lambda}_{ij,t'}}{\delta'_{ij,t}}.$$

Koutsopoulos and Xu (1993) claimed that actual information obtained on arc (i, j) will be less useful, as either the distance between the origin node and node i increases or the variability of the historical travel time on (i, j) increases. The new estimation of travel time (after being temporally and spatially discounted) on arc (i, j) is

$$\lambda^*_{ij,t_0+P_{si}(t_0)} = \bar{\lambda}_{ij,t''} + \mathrm{e}^{-\theta\sigma_{ij,t''}P_{si}(t_0)}(\lambda'_{ij,t''} - \bar{\lambda}_{ij,t''}),$$

where $P_{si}(t_0)$ is the shortest travel time from the origin node s to node i departing from the origin at time t_0, θ is a positive constant scalar that can be adjusted to produce a good fit between the estimated and actual travel times, $t''-\Delta \leqslant t_0+P_{ij}(t_0) \leqslant t''$, and $\sigma_{ij,t''}$ is the standard deviation of historical travel time $\bar{\lambda}_{ij,t''}$. The larger the value of $P_{si}(t_0)$ and $\sigma_{ij,t''}$, the larger the discounting of the actual information. This travel time updating procedure is incorporated in the label setting algorithm to find the shortest routes from an origin s. For each arc (i, j) out of the last permanently labeled node i, calculate (if node j is not yet permanently labeled):

$$P_{sj}(t_0) = \min\bigl\{P_{sj}(t_0), P_{si}(t_0) + \bar{\lambda}_{ij,t''} + \mathrm{e}^{-\theta\sigma_{ij,t''}P_{si}(t_0)}(\lambda'_{ij,t''} - \bar{\lambda}_{ij,t''})\bigr\}.$$

Set the label of a node with the smallest $P_{sj}(t_0)$ to permanent and update its predecessor node, which is needed to construct a path from the origin. The process is repeated until all nodes are labeled permanently.

Yang (2001) discussed an adaptive route selection with real-time updates in discrete STV networks, which is applied to ITS. To update the travel times, Yang considered both spatial and temporal information discounting, which are determined by spatial and temporal depth. The spatial depth determines the maximum reachable distance, with respect to the number of arcs, from the current node. The temporal depth is defined as the maximum number of time periods in which the information is still considered valuable. Furthermore, Yang also proposed two re-optimization algorithms to find the new adaptive route strategy that incorporates the new estimated travel times. The re-optimization algorithms are based on the ELB (Expected Lower Bound) algorithm of Miller-Hooks and Mahmassani (2000) and the SDOT algorithm of Miller-Hooks (2001). These re-optimization algorithms, called "adapted ELB" and "adapted SDOT," assume that the realization of the travel time must coincide with one of the possible values known a priori. Hence, it is assumed that the analysts are able to predict all possible values of future travel times, which is not realistic in many cases.

4.4 Global routing problems

The global route planning problem typically belongs to a government agency charged with the management of hazmat shipments within and through its jurisdiction. Although the transportation industry has been deregulated in many countries, hazmat transportation usually remains as part of the governments' mandate mainly due to the associated public and environmental risks. The two main concerns for a government agency are the *total risk* and the *spatial distribution of risk* in its jurisdiction. A number of policy tools are available to the government in mitigating public risk. These include proactive measures such as the establishment of inspection stations (Gendreau et al., 2000), insurance requirements (Verter and Erkut, 1997), and container specifications (Barkan

et al., 2000) as well as reactive measures such as the establishment of hazmat emergency response networks (Berman et al., 2007). Another common tool for governments is to ban the use of certain road segments by potentially hazardous vehicles. For an example of such regulation, we refer the reader to the local authority bylaws section of the Alberta Dangerous Goods Transportation and Handling Act (Government of Alberta, 2002). In the context of global route planning, the road segments to be closed by the government can be identified by solving a hazmat network design problem, which we discuss in Section 4.4.2.

Equity in the spatial distribution of risk can be important for a government agency for two reasons: (i) the perception of risk inequity frequently results in public opposition to the routing of vehicles carrying hazmats through the nearby passageways; and (ii) the overloading of certain road segments with hazmat flows (i.e., risk inequity) may lead to an increase in the incident probabilities as well as the severity of consequences. The concept of equity has been studied in the OR literature primarily within the context of undesirable facility location. Marsh and Schilling (1994) provided a comprehensive review of equity measures for location problems. Erkut (1993) offered two equity axioms for location problems and showed that the Gini coefficient and the coefficient of variation are the only measures that satisfy both of these axioms. Defining $n =$ number of zones, $t_i =$ individual risk at population zone i, and $\bar{t} =$ average individual risk, these two equity measures can be represented as follows:

$$\text{Coefficient of variation} = \frac{\sqrt{\sum_i (t_i - \bar{t})^2}}{n\bar{t}},$$

$$\text{Gini coefficient} = \frac{\sum_j \sum_i |t_i - t_j|}{2n^2\bar{t}}.$$

Coefficient of variation evaluates equity in terms of the deviations of the individual risks from the average. In contrast, Gini coefficient focuses on the differences between individual risks. Clearly, smaller values of these equity measures correspond to higher levels of fairness in risk distribution. A value of zero represents perfect equity, whereas a value of one represents absolute inequality. Using GIS, Verter and Kara (2001) estimated these two equity measures for gasoline shipments in Ontario and Quebec under four routing criterion: minimum length, minimum expected risk, minimum population exposure, and minimum incident probability.

4.4.1 Equity considerations in global route planning

The multiobjective model proposed by Zografos and Davis (1989) was perhaps the first attempt to explicitly incorporate equity considerations in global route planning for dangerous goods shipments. Their objectives were to minimize the total risk, the risk imposed on special population categories, travel time, and property damage. Equity is achieved by constraining the capacity of the road links. Zografos and Davis used pre-emptive goal programming in

solving the problem, and demonstrated (using hypothetical data) that forcing equity could increase the total risk up to 35%.

Gopalan et al. (1990a) proposed an equity constrained shortest path model that minimizes the total risk of travel between and origin–destination pair, while maintaining a desired level of equity among disjoint zones of a transportation network. Each zone constitutes a jurisdiction of a government agency that regulates hazmat transportation. The travel risk associated with road link (i, j) is the sum of risks imposed on the zones in the vicinity of the link. An origin–destination path is considered equitable if the difference between the risks imposed on any two arbitrary zones is under a given threshold. This equitable path definition can be incorporated in the shortest path model through additional constraints. Gopalan et al. (1990b) developed a subgradient algorithm to solve the Lagrangian dual, which is obtained by relaxing the equity constraints. They proposed a labeling shortest path procedure to close any remaining duality gap. The model was applied to a 50-node network from Albany, New York.

Gopalan et al. (1990b) extended their earlier work so as to identify a set of routes to be utilized for T trips between a single and origin–destination pair. In this case the equity threshold for a zone pair is the sum of the risk differences over T trips. Note that the T routes do not need to be distinct in their model. Gopalan et al. (1990b) proposed a heuristic procedure that repeatedly solves single trip problems using a Lagrangian dual approach with the gap-closing procedure, as in Gopalan et al. (1990a). To avoid having T identical routes, the link risks are modified using information from the previous t routes during iteration $(t + 1)$. This iterative procedure can easily be adapted to multiple origin–destination pairs.

In extending Gopalan et al. (1990b), Lindner-Dutton et al. (1991) focused on finding an equitable sequence of T trips, where the cumulative risk incurred by any zone after $t < T$ trips is equitable to that incurred by the other zones in the previous t trips. Both integer programming and dynamic programming (DP) formulations of this problem were presented. Lindner-Dutton et al. (1991) showed that a DP approach combined with the relaxation and fathoming methods of the Branch and Bound algorithm (as described in Morin and Marsten, 1976) could not solve moderate size problems to optimality within reasonable time. Therefore, they developed five upper bound heuristics to tackle large problems.

Marianov and ReVelle (1998) proposed a linear optimization model to solve the global route planning problem that minimizes both total cost and (the exact version of) accident probability. To introduce equity, they used an upper bound on the total risk associated with each arc. Similarly, Iakovou et al. (1999) incorporated equity through the use of a capacitated transport network model. Their multicommodity network flow model has two objectives: minimize transport cost and minimize expected risk cost. They used a weighted sum of these costs in conducting a trade-off analysis. A two-phase solution procedure, simi-

lar to that of Gopalan et al. (1990a), was proposed. The model was applied to marine transportation of oil products in the Gulf of Mexico.

The studies on generation of a set of spatially dissimilar (not necessarily disjoint) paths are also relevant to equity considerations in global route planning (e.g., Akgün et al., 2000; Dell'Olmo et al., 2005). Iterative penalty method (IPM), gateway shortest paths method, and minimax method are among the procedures that can be used to generate such a set of paths set between an origin–destination pair. However, Akgün et al. (2000) showed that the gateway shortest path method may not be suitable for generating dissimilar paths. They posed the dissimilarity problem as a p-dispersion problem (Erkut, 1990). In the p-dispersion context, p of m candidate paths are selected so that the minimum spatial dissimilarity between any pair of selected paths is maximized. The m candidate paths can be constructed using k-shortest path method or IPM.

Erkut and Verter (1998) proposed four indexes to measure the dissimilarity among paths P_1 and P_2:

- Arithmetic average of two ratios:

$$1 - \frac{L(P_1 \cap P_2)}{2L(P_1)} + \frac{L(P_1 \cap P_2)}{2L(P_2)};$$

- Geometric average of two ratios:

$$1 - \sqrt{\frac{L(P_1 \cap P_2)^2}{L(P_1)L(P_2)}};$$

- Ratio of the intersection length and the length of the longest path:

$$1 - \frac{L(P_1 \cap P_2)}{\max\{L(P_1), L(P_2)\}};$$

- Ratio of the intersection length and the length of the union of the two paths:

$$1 - \frac{L(P_1 \cap P_2)}{L(P_1 \cup P_2)};$$

where $L(P)$ denotes the length of path P.

Dell'Olmo et al. (2005) provided a multicriteria formulation of the dissimilar path problem. They used travel distance and transport risk as their criteria. After finding the Pareto-optimal set of paths, a buffer zone is constructed for each path in this set. This buffer zone approximates the impact area of a hazmat incident. Based on the buffer zones, a dissimilarity index can be calculated for each pair of paths by replacing $L(P)$ in the above definitions with $A(P)$ that represents the area of the buffer zone around path P. For example, the average arithmetic dissimilarity index can be defined as

$1 - A(P_1 \cap P_2)/(2A(P_1)) + A(P_1 \cap P_2)/(2A(P_2))$. A subset of maximally dissimilar paths (spatially speaking) can thus be found, for example, by applying the p-dispersion method.

The above models can be useful in identifying a global routing plan for a major hazmat producer/carrier that takes into account the equitable distribution of transport risk in a region. However, these models are of little use in the implementation of a comprehensive global transportation plan in a jurisdiction with multiple carriers since governments have no authority to impose routes on individual carriers. Yet many governments have the authority to close certain road segments to hazmat shipments (permanently or during certain hours of the day), and equity concerns can be incorporated into a hazmat network design problem. This is an interesting and challenging OR problem that has not been studied in the past. In the next section we review a closely related problem: the hazmat network design problem with a risk minimization objective.

4.4.2 Hazmat transportation network design

Network design problems have wide applications in both transportation and telecommunication planning (see, e.g., Magnanti and Wong, 1984; Balakrishnan et al., 1997). It is important to recognize the differentiating characteristics of this problem in the context of dangerous goods shipments. The transportation infrastructure is built mainly to connect heavily populated areas and not to avoid them. Therefore, the question becomes which road segments to close in an existing network rather than identifying the most appropriate ways to expand the infrastructure. Kara and Verter (2004) provide the following definition: given an existing road network, the hazardous network design problem involves selecting the road segments that should be closed to hazmat transport so as to minimize total risk. The carriers will select minimum cost routes on the designated hazmat network, and they are likely to incur higher costs due to reduced availability of routes. Hence, this can be considered a two-level decision problem where the government designates a subset of the transport network for hazmat transport and carriers select routes on this subset.

Note that these two levels cannot be considered in isolation. If one were to select minimum risk routes and offer the union of such routes to the carriers, the carriers would select minimum cost routes on this network which could result in much higher risk levels than the government had intended. This can be illustrated using the example depicted in Figure 21(a) (Erkut and Gzara, 2005). Suppose that hazmat type 1 is to be sent from node 1 to node 8, and hazmat type 2 is to be sent from node 2 to node 8. Assume that the transport cost for each commodity is the same.

If the carrier is allowed to route freely, it will select the minimum cost routes $\{(1,3),(3,8)\}$ and $\{(2,5),(5,6),(6,8)\}$ with a total cost of $3+3=6$ units and total risk of $8+8=16$ units. In contrast the minimum risk routes are $\{(1,3),(3,6),(6,8)\}$ and $\{(2,5),(5,6),(6,7),(7,8)\}$ with a total risk of

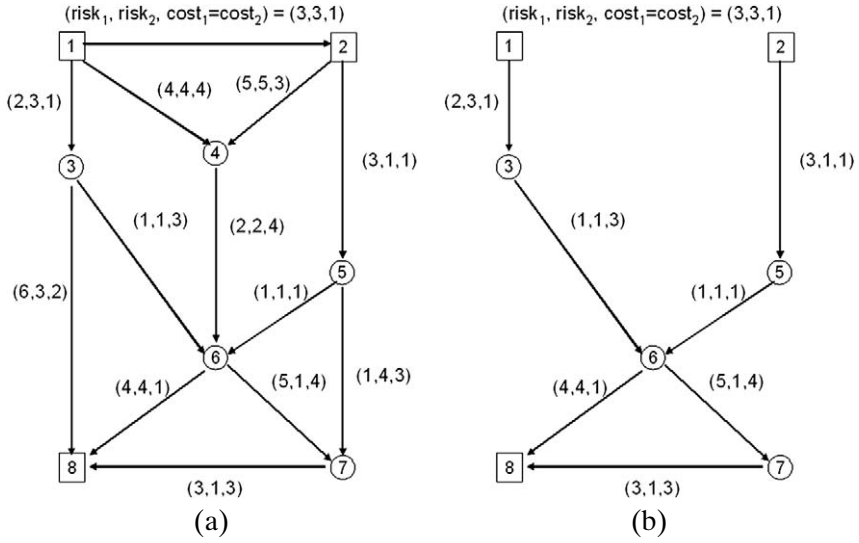


Fig. 21. (a) Multicommodity hazmat transport network design problem; (b) The feasible roads for routing cost minimization.

$7 + 4 = 11$ units. Figure 21(b) shows the union of the two minimum risk paths. If the government designates this network as the hazmat transport network, but allows the carrier to choose its routes, it will select the minimum cost routes $\{(1, 3), (3, 6), (6, 8)\}$ and $\{(2, 5), (5, 6), (6, 8)\}$ with a total cost of 8 units and total risk of 13 units. This risk is higher than what the government anticipates. As this example demonstrates, the design problem cannot be simplified to a one-level risk minimization problem, and the government must take into account the cost-minimizing behavior of carriers in designing the network.

The hazmat transportation network design problem has received the attention of researchers only recently. Kara and Verter (2004) proposed a bi-level integer linear programming formulation for this design problem that involves multiple types of hazmats. Their aim is to design a transport network so that the total risk resulting from the carriers' route choices is minimized. At the outer-level, risk is measured as the total number of people exposed to hazmat transport incidents. The inner-level problem represents the carriers' routing decisions on the available transport network so as to minimize their cost. This problem is represented by the linearized *Karush–Kuhn–Tucker* (KKT) conditions of its LP relaxation. As a result, the bi-level integer programming (IP) problem is transformed into a single-level mixed integer programming problem. The proposed model is solved by using CPLEX and applied to the hazmat transport network in Western Ontario, Canada. Kara and Verter demonstrate that carriers can benefit from the government's efforts and involvement in the regulation of dangerous goods shipments.

Erkut and Gzara (2005) considered a bi-level bi-objective (cost and risk minimization) network design problem similar to that discussed by Kara and Verter (2004). They proposed a heuristic algorithm that exploits the network flow structure at both levels, instead of transforming the bi-level IP problem to a single-level formulation. As a result, they achieved a significant increase in the computational performance.

Erkut and Alp (2007) posed the minimum risk hazmat network design problem as a Steiner tree selection problem. This topology takes away the carriers' freedom in route selection and simplifies the bi-level problem to a single level. However, it also results in circuitous (and expensive) routes. To avoid an economically infeasible solution, they suggested adding edges to the Steiner tree. They proposed a greedy heuristic that adds shortest paths to the tree so as to keep the risk increase to a minimum. They also posed a bi-objective version of the problem to minimize cost and risk, and solved it using a weighted additive objective. Their approach allows the decision maker to determine the density of the hazmat network where the options range from a tree to a completely connected network.

Verter and Kara (2005) provided a path-based formulation for the hazardous network design problem. Their main modeling construct is a set of alternative paths for each shipment. This facilitates the incorporation of carriers' cost concerns in regulator's risk reduction decisions. Paths not economically viable for carriers can be left out of the model. Alternative solutions to the network design problem can be generated by varying the number of routing options included in the model. To this end, Verter and Kara use pre-specified thresholds, e.g., for the maximum acceptable additional travel distance compared to the shortest path. Therefore, each solution corresponds to a certain compromise between the regulator and the carriers in terms of the associated transport risks and costs. Information about the nature of the cost-risk trade off can facilitate negotiation between the two parties. By using a GIS-based model of Quebec and Ontario, the authors demonstrate that their path-based formulation can be used for identifying road closure decisions that are mutually acceptable.

5 Facility location and transportation

Hazmat shipments often originate from facilities that themselves are potentially harmful to public and environmental safety, such as petroleum refineries or nuclear power plants. Also, the destinations of hazmat shipments can be noxious facilities such as gas stations and hazardous waste treatment centers. The location decisions pertaining to such facilities have a considerable effect on the routing of hazmat shipments. Therefore, integration of facility location and routing decisions can be an effective means to mitigate the total risk in a region where hazmats are processed and transported. It is interesting to

note that, in general, location decisions are considered strategic, whereas routing decisions are dealt with at the tactical level. However, the risk constitutes a coupling factor for these decisions in the context of dangerous goods. We refer the reader to Erkut and Neuman (1989) and Cappanera (1999) for extensive surveys of the location-only literature dealing with undesirable facilities. In this section, we provide a review of the prevailing studies on integrated location and routing models for hazmats.

The *location–routing problem* (LRP) involves determining the optimal number, capacity, and location of facilities as well as the associated optimal set of routes (and shipping schedules) to be used in serving customers. The distribution of goods from the facilities to the customers can be on a full-truck load or less than full-truck load basis. In the latter case, routes involving multiple customers are commonly used. From the solution method perspective, the LRP is NP-hard and offers a variety of challenges to OR researchers. The literature addressing LRP with different real-world applications has evolved since the late 1960s. Christofides and Elon (1969) were among the first to consider LRP with multiple customers on each route. The literature surveys on LRP include Madsen (1983), Balakrishnan et al. (1997), and Min et al. (1998).

Two types of risk need to be taken into account in integrating location and routing decisions pertaining to hazmat shipments: transport risk, R^{T}, and facility risk, R^{F}. Figure 22 illustrates these two types of risk. An individual at point x is exposed to (i) a transport incident on a nearby route segment l of a path P that involves a vehicle carrying volume v_P and (ii) an incident at the hazmat treatment center at site j with capacity u_j. The transport risk, $R^{\mathrm{T}}_{Pl}(v_P, x)$, can be determined as a function of the undesirable consequence at point x, taking into account the impact zone of a hazmat incident on segment l (see Section 3), and the estimated incident probability. The facility risk, $R^{\mathrm{F}}_j(u_j, x)$, can be determined in a similar way, with site j replacing the route segment l. Let $\mathbf{O}$ and $\mathbf{D}$ denote sets of origins and destinations, respectively, $\mathbf{P}_{OD}$ denote the set of all utilized paths for each O–D pair ($O \in \mathbf{O}$ and $D \in \mathbf{D}$), and $\mathbf{L}$ denote the set of hazmat facility locations. Assuming additivity of risk, the individual risk at point x can be determined as

$$R(x) := \sum_{O\in\mathbf{O}, D\in\mathbf{D}} \sum_{P\in\mathbf{P}_{OD}} \sum_{l\in P} R^{\mathbf{T}}_{Pl}(v_P, x) + \sum_{j\in\mathbf{L}} R^{\mathrm{F}}_j(u_j, x).$$

Let A denote the region of interest and $POP(x)$ denote the population density at point $x \in A$. The total risk in A is

$$R(A) = \int_{x\in A} R(x)POP(x)\,\mathrm{d}x.$$

Now consider a location–routing problem where $\mathbf{L} = \mathbf{D}$ (e.g., storage locations for spent nuclear fuel shipments). Let V_O denote the hazmat volume at $O \in \mathbf{O}$ (e.g., a nuclear power plant) that needs to be transported, and let u_D denote the capacity of a hazmat treatment facility at site $D \in \mathbf{D}$. Note that

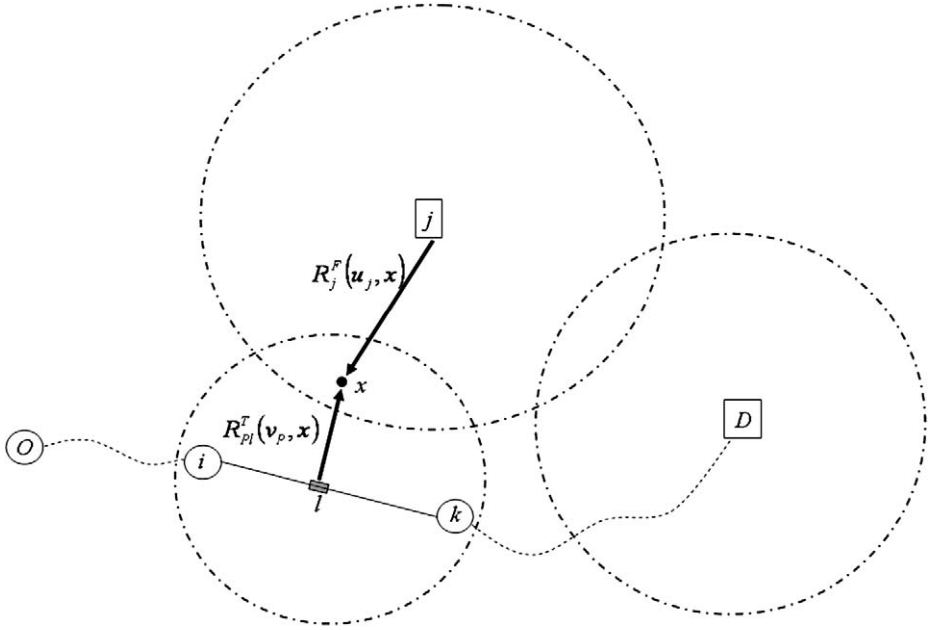


Fig. 22. Individual risk at point x due to transportation and processing of dangerous goods (adapted from List and Mirchandani, 1991).

D and $\mathbf{P}_{OD}$ now represent the sets of candidate locations for hazmat treatment facilities and the set of potential paths for each origin–destination pair, respectively. The set $\mathbf{P}_{OD}$ may represent the set of available routes on the hazmat road network designated by the government (see Section 4.4.2). We define two types of variables:

- binary location variables y_D, where

$$y_D = \begin{cases} 0, & \text{if a new hazmat treatment facility is located} \\ & \text{in site } D, \\ 1, & \text{otherwise,} \end{cases}$$

- nonnegative continuous flow variables v_P representing the quantity of hazmat shipped along path P.

Thus, the total risk in region A is

$$R(A) := \int_{x \in A} \Bigg(\sum_{O \in \mathbf{O}, D \in \mathbf{D}} \sum_{P \in \mathbf{P}_{OD}} \sum_{l \in P} R_{Pl}^{\mathrm{T}}(v_P, x) + \sum_{D \in \mathbf{D}} R_D^{\mathrm{F}}(u_D, x) y_D \Bigg) POP(x)\, \mathrm{d}x.$$

In addition to the total risk, the costs (i.e., transportation, operation, and fixed costs) should be also minimized. Let c_P^{T} denote the transportation cost per unit volume of hazmat along path P, c_D^{F} denote the (annualized) installation cost and c_D^{O} denote the unit operation cost of a hazmat treatment facility at site D. The total cost, TC, is determined as

$$TC := \sum_{O\in\mathbf{O},D\in\mathbf{D}} \sum_{P\in\mathbf{P}_{OD}} c_P^{\mathrm{T}} v_P + \sum_{D\in\mathbf{D}} \left(c_D^{\mathrm{F}} y_D + c_D^{\mathrm{O}} \sum_{O\in\mathbf{O}} \sum_{P\in\mathbf{P}_{OD}} v_P \right).$$

Also, equity in the spatial distribution of risk due to the location and routing decisions can be a relevant objective. Risk equity can be enforced, for example, by minimizing the maximum individual risk in the region, i.e.,

$$\overline{R}(A) := \max_{x\in A} R(x).$$

Hence, a mathematical programming formulation of the capacitated LRP to minimize the total risk and total cost and to force the risk equity can be constructed as follows:

$$\min \; R(A) \tag{5.1}$$

$$TC \tag{5.2}$$

$$\overline{R}(A) \tag{5.3}$$

subject to:

$$\sum_{D\in\mathbf{D}} \sum_{P\in\mathbf{P}_{OD}} v_P = V_O, \quad \text{for all } O \in \mathbf{O}, \tag{5.4}$$

$$\sum_{O\in\mathbf{O}} \sum_{P\in\mathbf{P}_{OD}} v_P \leqslant u_D y_D, \quad \text{for all } D \in \mathbf{D}, \tag{5.5}$$

$$\overline{R}(A) \geqslant \sum_{O\in\mathbf{O},D\in\mathbf{D}} \sum_{P\in\mathbf{P}_{OD}} \sum_{l\in P} R_{Pl}^{\mathrm{T}}(v_P, x) + \sum_{D\in\mathbf{D}} R_D^{\mathrm{F}}(u_D, x) y_D, \quad \text{for all } x \in A, \tag{5.6}$$

$$y_D \in \{0, 1\}, \quad \text{for all } D \in \mathbf{D}, \tag{5.7}$$

$$v_P \geqslant 0, \quad \text{for all } P \in \mathbf{P}_{OD} \text{ and } O\text{–}D \text{ pairs}, \; O \in \mathbf{O}, D \in \mathbf{D}. \tag{5.8}$$

Constraints (5.4) ensure that all hazmat generated must be shipped out of the origins, whereas constraints (5.5) stipulate that if a facility at location D is open (i.e., $y_D = 1$), then total quantity of hazmat to be treated at D cannot exceed the pre-specified capacity of the facility. Constraints (5.6) are used to incorporate the risk equity. It is evident from the above model that the hazmat LRP is multiobjective by nature. The surveys by List et al. (1991), Boffey and Karkazis (1993), and Cappanera et al. (2004) observed that literature on hazmat LRP is

sparse. In this section, rather than duplicating these surveys, we highlight the important results.

Shobrys (1981) is the first study on hazmat LRP with a focus on selecting routes and storage locations for spent nuclear fuel shipments. A decomposition approach is used to separate the routing problem from the location problem. Two routing objectives are minimized; ton-miles and population exposure-tons. The associated bi-objective shortest path model identifies a set of Pareto-optimal paths between each waste source (origin) and each candidate storage site (destination). The weighted costs associated with each Pareto-optimal path determine the cost coefficients of the p-median problem that is used to select the storage site.

Zografos and Samara (1989) considered an LRP with three objectives, namely minimization of transport risk, minimization of travel times, and minimization of disposal risk, to establish locations of a given number of waste treatment facilities and determine the associated shipment routes. Their model requires that the hazardous waste at each population center must be disposed of entirely. Each population center is assigned to its nearest disposal facility. Moreover, links of the transportation network are capacitated. Pre-emptive goal programming is used to generate solutions under a few different scenarios.

List and Mirchandani (1991) proposed a hazmat LRP model that simultaneously considers total transportation and treatment risk, total transportation cost, and risk equity. Risk equity is enforced by minimizing the maximum consequence per unit population for all mutually disjoint zones of the transportation network. Their formulation served as a basis for the model in (5.1)–(5.8). However, the List and Mirchandani model is more general since it allows for different types of hazardous materials and treatment technologies. This model assumes that the impact to point x in a zone Z from a vehicle incident is inversely proportional to the square of the Euclidean distance between the vehicle and point x, and the impact is directly proportional to the volume v_P being shipped regardless of material. Hence, the transport risk faced by an individual at point x is determined as

$$R^{\mathrm{T}}_{Pl}(v_P, x) := \alpha v_P \int_{l \in P} \|l - x\|^{-2} c(x) \pi(l)\, \mathrm{d}l,$$

where α is a constant of proportionality, $c(x)$ is a likelihood of impact at point x, and $\pi(l)$ is the probability of an incident at road segment l. The facility risk from an incident at a hazardous waste treatment facility at site j of waste type w with treatment technology t and volume u_{jwt}, $R^{\mathrm{F}}_{jwt}(u_{jwt}, x)$, is determined in a similar way. However, their facilities have unlimited capacity and the total cost of establishing treatment facilities is bounded by a budget constraint. Uncertainty is considered in constructing the risk formulations, but it is not incorporated in solving the example case. Instead, the expected number of fatalities is used to calculate the risk. The LRP problem is solved using LINDO. The weighted sum technique is used to study the tradeoffs among

the objectives in identifying the transportation routes, locating the hazardous waste treatment facilities, and choosing the treatment technologies.

ReVelle et al. (1991) developed a combined discrete location–routing model for shipments of spent nuclear fuel that minimizes both transportation cost and perceived risk. As in Shobrys (1981), the transportation cost is measured in ton-miles, and the perceived risk is measured using population exposure as people-tons. The total people-ton of an arc is the product of the number of people within a certain bandwidth on the arc and the tons of hazardous waste shipped on that arc. The problem is solved in two stages. In the first stage, a weighted sum of the arc distance and the number of people in the impact area around that arc (called *hybrid distance*) is calculated for every arc in the network. Floyd's shortest path algorithm is used to generate (hybrid) shortest paths for all origin–destination pairs. In the second stage, the location problem is modeled as a p-median problem, where the coefficients of the objective function are calculated by taking the product of the tons of spent fuel at the origin and the hybrid shortest distance from the origin to the destination.

Stowers and Palekar (1993) proposed a bi-objective network LRP with a single facility and a single commodity. In a network LRP, the waste facility can be located anywhere on the network. Two objectives are considered, namely minimizing the total exposure (minisum) and minimizing the maximum exposure (minimax). The total exposure to a node or to an arc of the network is represented as a convex combination of location exposure and travel exposure, where the impact area is modeled as a danger circle. Stowers and Palekar showed that an optimal solution to the minisum and minimax problems with only travel exposure occurs at a node. The nodal optimality is still valid for any positive linear combination of travel cost and travel exposure as long as the travel cost is an increasing function of distance, as in ReVelle et al. (1991). Moreover, when population is concentrated at nodes only, a finite dominating set of facility locations can be identified which is guaranteed to contain an optimal solution.

Giannikos (1998) proposed a multiobjective model for a discrete hazardous waste LRP that minimizes the following four objectives:

(1) total transportation cost and fixed cost of opening the treatment facilities;
(2) total perceived risk due to the shipment of hazardous waste;
(3) maximum individual risk (to force the risk equity); and
(4) maximum individual disutility due to the treatment facilities.

The disutility imposed on a population center i by the establishment of a treatment facility at site j is a function of the capacity of facility j and the distance between i and j. The total disutility at population center i is obtained by adding the disutilities imposed upon i by all treatment facilities. A weighted goal programming technique is used to solve the problem.

Cappanera et al. (2004) presented a single objective LRP model that minimizes the total transportation and facility establishment costs. In their model,

an arc formulation is given instead of a path formulation as in (5.1)–(5.8). Their model includes constraints that require both routing and population exposures for each affected site to remain within given threshold values. Arcs of the network are incapacitated, but the facilities are capacitated. Cappanera et al. (2004) consider only a single commodity and seek to find the optimal number of facilities. By dualizing the capacity constraints, the LRP is decomposed into location and routing subproblems to obtain a lower bound. To find the upper bounds, two *Lagrangian heuristics*, called the *Location–Routing heuristic* and *Routing–Location heuristic*, are proposed.

In closing this section, we note that almost all existing models for hazmat LRP are static and deterministic. Only the model of List and Mirchandani (1991) considers different types of hazmats and technology selection for hazmat treatment facilities as well as uncertainty in problem parameters. The lack of multiple hazmat models that consider stochasticity in a time-dependent environment constitutes an area for further LRP research.

6 Synthesis and future research directions

To summarize the material we have reviewed, Tables 2(a–d) groups the models into classes distinguished by

- the main aspects of the problem (risk assessment, routing, combined facility location and routing, and network design),
- transport mode,
- single vs. multiple objectives,
- whether or not stochastic elements are included,
- whether or not time-variant elements are included,
- whether or not GIS is used.

Tables 2(a–d) suggests that the hazmat transportation problems on highways received the most attention from the operations researchers. In contrast, hazmat transportation via air or pipeline, as well as intermodal hazmat transportation has received almost no attention. From the methodological perspectives, we observe that:

- global routing problems on stochastic time-varying networks received no attention despite their relevance and application potential,
- hazmat transportation network design problem which considers all involved parties (government and the carriers) is a relatively young research topic. The most obvious extension of the existing models in this area is to incorporate uncertainty and consider multiple objectives as the hazmat transportation problems are highly stochastic in nature and involve multiple criteria (and players),
- there is an increase on utilizing a GIS either for data input or combined with optimization models to conduct more realistic risk assessment.

Erkut and Verter (1995a) reflected on the state-of-the-art as of 1995, and pointed out a number of directions for future research. In the following ten years, some of the problem areas proposed in Erkut and Verter (1995a) were investigated by researchers, whereas many others remained relatively unexplored. We discuss some of the underexploited areas discussed in Erkut and Verter (1995a), as well as other potential problem areas, that can lead to fruitful research.

Risk calculation – probabilities

QRA relies heavily on empirical accident/incident probabilities. However past data is not very reliable. Using general truck accident data for hazmat trucks overestimates the accident probabilities. What makes matters worse is that there is no agreement on general truck accident probabilities and conflicting numbers are reported by different researchers. Furthermore, applying national data uniformly on all road segments of similar type is quite problematic since it ignores hot spots such as road intersections, highway ramps, and bridges. Researchers need to have access to high quality accident probability data and empirical or theoretical research that leads to improvements in the quality of such data would be welcome.

Risk calculation – perceived risks

Given the limitation of QRA, and the fact that public opposition is a function of perceived risks, perhaps more attention should be paid to quantifying and modeling of perceived risks. We believe more work is needed to improve our understanding of how perceived risks change as a function of the hazardous substance, the distance to a hazardous activity, and the volume of the activity.

Risk calculation – consequences

The second important input in QRA is the population exposed as a result of an incident. Many past studies used uniform population density along transport links which is a very blunt approach. A GIS makes it possible to use more precise population information. However, using census-based population data for daytime hazmat movements makes little sense since census data is residence-based and most residents are not at home during the day. Researchers need to take the next step and incorporate day versus night population distributions, as well as high-density population installations such as schools and hospitals. While this is done relatively easily for QRA of a single route, it is more complicated to generate the necessary data for an entire transportation network.

Risk calculation – time dependence

There are very significant differences in risks between day and night (due to differences in accident probabilities, population distributions, and weather conditions). Yet most of the OR literature pays little attention to this. Risk radii (or safe distances) strongly depend on transport mode and weather conditions. Hence, it is impossible to speak of a single "minimum risk" route; hazmat routing problems must be solved with real-time information. Solving problems with static parameter values can result in poor solutions and decisions.

Risk calculation – model

We emphasized the importance of using the proper risk model throughout the chapter. It is important to use as accurate a model as technically and computationally feasible. For example, it is not only possible, but also necessary to combine GIS data, plume dispersion modeling, and real-time weather information to determine bypass routes for chlorine shipments. In fact, analysis that does not use such level of detail is of little use in the case of hazmats that can generate plumes.

Risk calculation – nonhuman risks

The vast portion of the hazmat risk literature is concerned with fatalities, and to some extent injuries and property damage. Little if any attention is paid to environmental damage. Environmental risks are usually only included in multiattribute utility models. We believe that hazmat risk models should take into account all risks to humans and environment for broader acceptance by the public.

Multicriteria approach to risk minimization

It is well known that different routes can emerge as minimum risk routes depending on the definition of risk used. Hence, it is crucial to use multiple measures and provide decision-makers with a set of efficient solutions instead of a single "risk minimizing" route. Development of methodology that would allow for the decision-makers to effectively search the efficient solution set and select a route would be of great practical use.

Risk equity

The academic literature suggests that equity in the spatial distribution of risk is a critical concern in designing hazmat management strategies acceptable to the public. Yet, risk equity is not a great concern to the hazmat industry. If equity is a valid concern then it must be imposed by a regulatory agency.

Local vs. global route planning

Most hazmat transport models deal with only one commodity. While it may make sense for carriers to decompose a transport planning problem into multiple single commodity problems, if one is concerned about concentration and distribution of risks, one has to pose a multicommodity problem where risk and equity concerns couple the different materials. For example, hazmat facility location models should include the hazmat distribution network for proper risk assessment. Likewise, the hazmat network design problem requires consideration of all hazmats.

Multidisciplinary nature of the problem

It is rather unfortunate that research in this highly multidisciplinary area continues to be compartmentalized. Chemical and civil engineers tend to publish in their own journals, decision analysts and quantitative risk assessment researchers limit their focus to their paradigms, and operations researchers seldom wonder outside their safe zone. For fruitful research and applications researchers from different disciplines have to reach out to one another.

Cost consequences of risks

One of the reasons why hazmat carriers are not too interested in hazmat routing research is that there are no consequences to not using a decision-support system before making routing decisions. If carriers are faced with lawsuits as a result of poor routing decisions, or if their insurance companies (or creditors) required the use of QRA in route planning to avoid such lawsuits, or if a government agency required the use of QRA and OR tools in route planning, we believe that research in this area would accelerate considerably.

Implementation

It is inconceivable to imagine a hazmat transport DSS that does not take advantage of a GIS while most academic researchers solve small problems on made-up (*realistic*) networks. In fact the ideal hazmat transport DSS would combine GIS, QRA, OR, and MCDA. We suggest that research in this area follow the same recipe. This increases adoption probability by the industry. We note that clever use of GIS can enable one to incorporate nonhuman risks into the analysis. Another necessary condition for successful implementation of OR research in this area is cooperation between the researchers, the government agencies, and the carriers – something we cannot claim has happened with regularity in the past.

Recent concern: security

In addition to the concerns discussed above there is a new concern in hazmat transport planning, namely potential for a terrorist attack on a hazmat vehicle. The terrorist attacks in the USA in 2001 have focused attention on what other targets terrorists may choose. It was quickly recognized that hazmat vehicles could be desirable targets for terrorists, and certain hazmat vehicles were designated as "weapons of mass destruction" (TRB, 2002; Abkowitz, 2002). Such concerns changed the way the hazmat transport industry operates. For example, the US Federal Government now requires hazmat truckers to submit to fingerprinting and criminal background checks (Glaze, 2003).

This security issue, however, has not yet received much attention from operations researchers. Clearly, the problem is complex and there are many solutions that involve little or no OR. However, there is potential for OR contributions and we list three here:

- Rerouting around major cities: the risk of terrorist attacks made it very undesirable to route hazmat vehicles (particularly trains) through major population centers. Traditional OR algorithms can be used to find alternate routes for shipments. Erkut and Glickman (1997) show that significant risk reductions are possible through rerouting, and Erkut and Ingolfsson (2000) develop new methodology for routing with a catastrophe-avoidance objective.
- Changes in the modeling of incidence risks: The traditional risk assessment for hazmat transport assumes incidents are caused by traffic accidents or human error. Yet we now know that there is a nonzero probability of a terrorist attack or a hijack. Not only does this increase the incident probabilities, but it also requires a new way of modeling consequences since the impact may no longer be limited to the planned route. Furthermore, attack probabilities are unlikely to be uniform. For example, a location in a tunnel, on a bridge, or near a "trophy building" is likely to have a higher attack probability than a location in a remote and unpopulated area. In contrast, sparsely populated areas may be associated with a higher hijack probability. A hijacked vehicle's future route is unpredictable and special precautions may have to be taken to prevent it from having an incident in a highly populated area. As a result, traditional risk assessment-based route planning is no longer adequate. There are very few papers in this new area. (See Paté-Cornell, 2002, for probabilistic modeling of terrorist threats, and Huang and Cheu, 2004 and Huang et al., 2003, for incorporation of security concerns in route planning.)
- Changes in route planning methodology: Past hazmat routing literature focuses on finding a minimum risk route. The problem with determining quantitative measures and selecting routes accordingly is

that terrorists could predict such routes by using similar methods. To minimize the probability of a successful terrorist attack or hijacking, shippers could alternate routes – game theory can be applied to the problem of alternating among routes to minimize predictability – or change them en-route in real time in ways that would be difficult to predict. Video surveillance or Global Positioning Systems (GPS) and communication equipment installed on all hazmat vehicles would not only allow for precise tracking of vehicles, but also allow the implementation of such real-time decision making (see, e.g., Glaze, 2003; Zografos and Androutsopoulos, 2001).

We believe that there are still many important OR problems in hazmat transportation. However, we think the focus will shift from a priori optimization toward real-time adaptive decision making for several reasons, such as the availability of the necessary technology and data, as well as security concerns. While it is rather unfortunate that terrorist attacks can and do happen, their possibility opens up a new frontier for operations researchers in general, and hazmat transport researchers in particular. We expect that hazmat transport research will intensify in the near future and we hope that this chapter will be useful to future researchers in this area.

We finish with an attempt to explain why we find research in hazmat logistics particularly interesting and challenging, in addition to the potential for practical applications. The realm of OR can be crudely divided into two major paradigms: deterministic and stochastic. Optimization is the major tool in the deterministic area while the stochastic domain requires probabilistic modeling. Much of the research in OR can be classified in one of these two regions. Hazmat logistics research lies in the cross-section of these two domains, and it requires a good knowledge of probabilistic modeling as well as optimization techniques. Hazmat transport can be modeled as a probabilistic phenomenon, but one needs to add optimization of appropriate objectives to realize the possible policy benefits. The fact that we are modeling an inherently probabilistic process results in the natural consequence that there are many appropriate objectives. The exact probabilistic expressions are usually too complicated, which results in the use of approximations for optimization. Hence, the researchers must understand probabilistic modeling well enough to capture the essence of the activity, but they must also be sufficiently proficient in optimization techniques to decide which approximations are necessary and what tools to use. The multicriteria/multistakeholder nature of the problems adds to the complexity as well as the attraction of this area. We found research in hazmat logistics quite rewarding and we encourage others to explore this area further.

Acknowledgements

This research has been supported in part by two Discovery Grants from NSERC (OGP 25481 and 183631). The authors acknowledge the input pro-

vided by Manish Verma on rail and intermodal transportation of hazardous materials.

References

Abkowitz, M. (2002). Transportation risk management: A new paradigm. Security papers, Southeastern Transportation Center, University of Tennessee, pp. 93–103.

Abkowitz, M., Cheng, P.D.M. (1988). Developing a risk cost framework for routing truck movements of hazardous materials. *Accident Analysis and Prevention* 20 (1), 39–51.

Abkowitz, M., Cheng, P.D.M. (1989). Hazardous materials transport risk estimation under conditions of limited data availability. *Transportation Research Record* 1245, 14–22.

Abkowitz, M., Lepofsky, M., Cheng, P. (1992). Selecting criteria for designating hazardous materials highway routes. *Transportation Research Record* 1333, 30–35.

Abkowitz, M.D., DeLorenzo, J.P., Duych, R., Greenberg, A., McSweeney, T. (2001). Assessing the economic effect of incidents involving truck transport of hazardous materials. *Transportation Research Record* 1763, 125–129.

Akgün, V., Erkut, E., Batta, R. (2000). On finding dissimilar paths. *European Journal of Operational Research* 121 (2), 232–246.

Akgün, V., Parekh, A., Batta, R., Rump, C.M. (2007). Routing of a hazmat truck in the presence of weather systems. *Computers & Operations Research* 34 (5), 1351–1373.

ALK Associates (1994). ALK's PC*HazRoute (Version 2.0). ALK Associates, Inc., 1000 Herrontown Road, Princeton, NJ, USA.

Alp, E. (1995). Risk-based transportation planning practice overall methodology and a case example. *INFOR* 33 (1), 4–19.

Alp, E., Zelensky, M.J. (1996). Risk quantification for meteorology- and direction-dependent hazards due to point and linear risk sources. *Journal of Loss Prevention in the Process Industries* 9 (2), 135–145.

Alumur, S., Kara, B.Y. (2007). A new model for the hazardous waste location–routing problem. *Computers & Operations Research* 34 (5), 1406–1423.

American Institute of Chemical Engineers (2000). *Guidelines for Chemical Process Quantitative Risk Analysis*, 2nd edition. Center for Chemical Process Safety, New York.

Anderson, R.T., Barkan, C.P.L. (2004). Railroad accident rates for use in transportation risk analysis. *Transportation Research Record* 1863, 88–98.

Andersson, S.E. (1994). Safe transport of dangerous goods – road, rail or sea – a screening of technical and administrative factors. *European Journal of Operational Research* 75 (3), 499–507.

Ang, A., Briscoe, J. (1989). Development of a systems risk methodology for single and multimodal transportation systems. Final report, Office of University Research, US DOT, Washington, DC.

Arya, S.P. (1999). *Air Pollution Meteorology and Dispersion*. Oxford Univ. Press, New York.

Ashtakala, B., Eno, L.A. (1996). Minimum risk route model for hazardous materials. *Journal of Transportation Engineering – ASCE* 122 (5), 350–357.

Balakrishnan, A., Magnanti, T.L., Mirchandani, P. (1997). Network design. In: Dell'Amico, M., Maffioli, F., Martello, S. (Eds.), *Annotated Bibliographies in Combinatorial Optimization*. Wiley, Chichester.

Barkan, C.P.L., Treichel, T.T., Widell, G.W. (2000). Reducing hazardous materials releases from railroad tank car safety vents. *Transportation Research Record* 1707, 27–34.

Barkan, C.P.L., Dick, C.T., Anderson, R. (2003). Railroad derailment factors affecting hazardous materials transportation risk. *Transportation Research Record* 1825, 64–74.

Barnes, P. (2001). Regulating safety in an unsafe world (risk reduction for and with communities). *Journal Hazardous Materials* 86, 25–37.

Batta, R., Chiu, S.S. (1988). Optimal obnoxious paths on a network: Transportation of hazardous materials. *Operations Research* 36 (1), 84–92.

Belardo, S., Pipkin, J., Seagle, J.P. (1985). Information support for hazardous materials movement. *Journal Hazardous Materials* 10 (1), 13–32.

Bellman, R.E. (1958). On a routing problem. *Quarterly of Applied Mathematics* 16, 87–90.

Berman, O., Drezner, Z., Wesolowsky, G.O. (2000). Routing and location on a network with hazardous threats. *Journal of the Operational Research Society* 51 (9), 1093–1099.

Berman, O., Verter, V., Kara, B. (2007). Designing emergency response networks for hazardous materials transportation. *Computers & Operations Research* 34 (5), 1374–1388.

Beroggi, G.E.G. (1994). A real-time routing model for hazardous materials. *European Journal of Operational Research* 75 (3), 508–520.

Beroggi, G.E.G., Wallace, W.A. (1991). Closing the gap-transit control for hazardous material flow. *Journal of Hazardous Materials* 27 (1), 61–75.

Beroggi, G.E.G., Wallace, W.A. (1994). Operational risk management – a new paradigm for decision-making. *IEEE Transactions on Systems Man and Cybernetics* 24 (10), 1450–1457.

Beroggi, G.E.G., Wallace, W.A. (1995). Operational control of the transportation of hazardous materials: An assessment of alternative decision models. *Management Science* 41 (12), 1962–1977.

Boffey, T.B., Karkazis, J. (1993). Models and methods for location and routing decisions relating to hazardous materials. *Studies in Locational Analysis* 4, 149–166.

Boffey, T.B., Karkazis, J. (1995). Linear versus nonlinear models for hazardous materials routing. *INFOR* 33, 114–117.

Bonvicini, S., Leonelli, P., Spadoni, G. (1998). Risk analysis of hazardous materials transportation: Evaluating uncertainty by means of fuzzy logic. *Journal of Hazardous Materials* 62 (1), 59–74.

Bowler, L.A., Mahmassani, H.S. (1998). Routing of radioactive shipments in networks with time-varying costs and curfews. Technical Report ANRCP-1998-11, Amarillo National Resource Center for Plutonium, TX (United States).

Boykin, R.F., Freeman, R.A., Levary, R.R. (1984). Risk assessment in a chemical storage facility. *Management Science* 30 (4), 512–517.

Brown, D.F., Dunn, W.E. (2007). Application of a quantitative risk assessment method to emergency response planning. *Computers & Operations Research* 34 (5), 1243–1265.

Bubbico, R., Ferrari, C., Mazzarotta, B. (2000). Risk analysis of LPG transport by road and rail. *Journal of Loss Prevention in the Process Industries* 13 (1), 27–31.

CANUTEC (2004). *Emergency Response Guidebook*.

Cappanera, P. (1999). A survey on obnoxious facility location problems. Technical Report TR-99-11, University of Pisa.

Cappanera, P., Gallo, G., Maffioli, F. (2004). Discrete facility location and routing of obnoxious activities. *Discrete Applied Mathematics* 133, 3–28.

Carotenuto, P., Giordani, S., Ricciardelli, S. (2007a). Finding minimum and equitable risk routes for hazmat shipments. *Computers & Operations Research* 34 (5), 1304–1327.

Carotenuto, P., Giordani, S., Ricciardelli, S., Rismondo, S. (2007b). A tabu search approach for scheduling hazmat shipments. *Computers & Operations Research* 34 (5), 1328–1350.

Cassini, P. (1998). Road transportation of dangerous goods: Quantitive risk assessment and route comparison. *Journal of Hazardous Meterials* 61 (1–3), 133–138.

Chakraborty, J., Armstrong, M.P. (1995). Using geographic plume analysis to assess community vulnerability to hazardous accidents. *Computers, Environment, and Urban Systems* 19 (5/6), 341–356.

Chang, N.B., Wei, Y.L., Tseng, C.C., Kao, C.Y.J. (1997). The design of a GIS-based decision support system for chemical emergency preparedness and response in an urban environment. *Computers, Environment, and Urban Systems* 21 (1), 67–94.

Chang, T.S., Nozick, L.K., Turnquist, M.A. (2005). Multi-objective path finding in stochastic dynamic networks, with application to routing hazardous materials shipments. *Transportation Science* 39 (3), 383–399.

Chin, S.-M., Cheng, P.D.-M. (1989). Bicriterion routing scheme for nuclear spent fuel transportation. *Transportation Research Record* 1245, 60–64.

Chow, T.C., Oliver, R.M., Vignaux, G.A. (1990). A Bayesian escalation model to predict nuclear accidents and risk. *Operations Research* 38 (2), 265–277.

Christofides, N., Eilon, S. (1969). Expected distances in distribution problems. *Operational Research Quarterly* 20, 437–443.

Coleman, J.A. (1984). Railroad–highway crossings and route selection for transporting hazardous materials. *Public Roads* 48 (2), 63–71.

Corea, G., Kulkarni, V.G. (1993). Shortest paths in stochastic networks with arc lengths having discrete distributions. *Networks* 23, 175–183.

Cox, R., Turnquist, M. (1986). Scheduling truck shipments of hazardous materials in the presence of curfews. *Transportation Research Record* 1063, 21–26.

Current, J.R., Ratick, S. (1995). A model to assess risk, equity, and efficiency in facility location and transportation of hazardous materials. *Location Science* 3, 187–202.

Cutter, S.L., Ji, M.H. (1997). Trends in US hazardous materials transportation spills. *Professional Geographer* 49 (3), 318–331.

Dell'Olmo, P., Gentili, M., Scozzari, A. (2005). On finding dissimilar Pareto-optimal paths. *European Journal of Operational Research* 162, 70–82.

Deng, C., Oberg, S.G., Downs, J.L. (1996). Risk assessment for transportation of radioactive material within the state of Idaho. *Health Physics* 70 (6), 41. Annual Meeting of the Health Physics Society, Seattle, WA (United States).

Dennis, S.M. (1996). Estimating risk costs per unit of exposure for hazardous materials transported by rail. *Logistics and Transportation Review* 32 (4), 351–375.

Douligeris, C., Iakovou, E., Yudhbir, L. (1997). Maritime route risk analysis for hazardous materials transportation. *Proceedings of the 8th IFAC/IFIP/IFORS Symposium on Transportation Systems*, Chania, Crete, Greece, pp. 574–579.

Duque, J., Barbosa-Póvoa, A.P.F.D. (2007). Synthesis and optimization of the recovery route for residual products under uncertain product demand. *Computers & Operations Research* 34 (5), 1463–1490.

Efroymson, R.A., Murphy, D.L. (2000). Ecological risk assessment of multimedia hazardous air pollutants: Estimating exposure and effects. *Science of the Total Environment* 274, 219–230.

Erkut, E. (1990). The discrete p-dispersion problem. *European Journal of Operational Research* 46, 48–60.

Erkut, E. (1993). Inequality measures for location problems. *Location Science* 1 (3), 199–217.

Erkut, E. (1995). On the credibility of the conditional risk model for routing hazardous materials. *Operations Research Letters* 18, 49–52.

Erkut, E. (1996). The road not taken. *OR/MS Today* 23, 22–28.

Erkut, E., Alp, O. (2006). Integrated routing and scheduling of hazmat trucks with stops en-route. *Transportation Science*, in press.

Erkut, E., Alp, O. (2007). Designing a road network for dangerous goods shipments. *Computers & Operations Research* 34 (5), 1389–1405.

Erkut, E., Glickman, T. (1997). Minimax population exposure in routing highway shipments of hazardous materials. *Transportation Research Record* 1602, 93–100.

Erkut, E., Gzara, F. (2005). A bi-level programming application to hazardous material transportation network design. Research report, Department of Finance and Management Science, University of Alberta School of Business, Edmonton, Alberta, Canada.

Erkut, E., Ingolfsson, A. (2000). Catastrophe avoidance models for hazardous materials route planning. *Transportation Science* 34 (2), 165–179.

Erkut, E., Ingolfsson, A. (2005). Transport risk models for hazardous materials: Revisited. *Operations Research Letters* 33 (1), 81–89.

Erkut, E., Neuman, S. (1989). Analytical models for locating undesirable facilities. *European Journal of Operational Research* 40, 275–291.

Erkut, E., Verter, V. (1995a). Hazardous materials logistics. In: Drezner, Z. (Ed.), *Facility Location: A Survey of Applications and Methods*. Springer-Verlag, New York, pp. 467–506. Chapter 20.

Erkut, E., Verter, V. (1995b). A framework for hazardous materials transport risk assessment. *Risk Analysis* 15 (5), 589–601.

Erkut, E., Verter, V. (1998). Modeling of transport risk for hazardous materials. *Operations Research* 46 (5), 625–642.

Ertugrul, A. (1995). Risk-based transportation – planning practice – overall methodology and a case example. *INFOR* 33 (1), 4–19.

Fabiano, B., Curro, F., Palazzi, E., Pastorino, R. (2002). A framework for risk assessment and decision-making strategies in dangerous good transportation. *Journal of Hazardous Materials* 93 (1), 1–15.

Ferrada, J.J., Michelhaugh, R.D. (1994). Development of an expert system for transportation of hazardous and radioactive materials. In: *International Topical Meeting on Nuclear and Hazardous Waste Management, Spectrum '94, Atlanta*. American Nuclear Society, Inc., l.2, pp. 997–1002.

FMCSA (2001). Comparative risks of hazardous materials and non-hazardous materials truck shipment accidents/incidents. Federal Motor Carrier Safety Administration, Washington, DC.

Frank, H. (1969). Shortest paths in probabilistic graphs. *Operations Research* 17, 583–599.

Frank, W.C., Thill, J.C., Batta, R. (2000). Spatial decision support system for hazardous material truck routing. *Transportation Research Part C – Emerging Technologies* 8 (1–6), 337–359.

Fronczak, R.E. (2001). Public health risks of railroad hazardous substance emergency events. Letters to the Editor. *Journal of Occupational and Environmental Medicine* 43 (9), 738–739.

Fu, L., Rilett, L.R. (1997). Real-time estimation of incident delay in dynamic and stochastic networks. *Transportation Research Record* 1603, 99–105.

Fu, L., Rilett, L.R. (1998). Expected shortest paths in dynamic and stochastic traffic networks. *Transportation Research Part B* 32 (7), 499–516.

Gendreau, M., Laporte, G., Parent, I. (2000). Heuristics for the location of inspection stations on a network. *Naval Research Logistics* 47 (4), 287–303.

Giannikos, I. (1998). A multiobjective programming model for locating treatment sites and routing hazardous wastes. *European Journal of Operational Research* 104 (2), 333–342.

Glaze, M. (2003). New security requirements for hazmat transportation. *Occupational Health & Safety* 72 (9), 182–185.

Glickman, T.S. (1983). Rerouting railroad shipments of hazardous materials to avoid populated areas. *Accident Analysis and Prevention* 15 (5), 329–335.

Glickman, T.S. (1988). Benchmark estimates of release accident rates in hazardous materials transportation of rail and truck. *Transportation Research Record* 1193, 22–28.

Glickman, T.S. (1991). An expeditious risk assessment of the highway transportation of flammable liquids in bulk. *Transportation Science* 25 (2), 115–123.

Glickman, T.S., Golding, D. (1991). For a few dollars more: Public trust and the case for transporting nuclear waste in dedicated trains. *Policy Studies Review* 10, 4.

Glickman, T.S., Rosenfield, D.B. (1984). Risks of catastrophic derailments involving the release of hazardous materials. *Management Science* 30 (4), 503–511.

Glickman, T.S., Sontag, M.A. (1995). The tradeoffs associated with rerouting highway shipments of hazardous materials to minimize risk. *Risk Analysis* 15 (1), 61–67.

Gopalan, R., Kolluri, K.S., Batta, R., Karwan, M.H. (1990a). Modeling equity of risk in the transportation of hazardous materials. *Operations Research* 38 (6), 961–975.

Gopalan, R., Batta, R., Karwan, M.H. (1990b). The equity constrained shortest path problem. *Computers & Operations Research* 17 (3), 297–307.

Government of Alberta (2002). Dangerous goods transportation and handling act. Available at *http://www.qp.gov.ab.ca/documents/acts/D04.cfm*.

Gregory, R., Lichtenstein, S. (1994). A hint of risk-tradeoffs between quantitative and qualitative risk-factors. *Risk Analysis* 14 (2), 199–206.

Grenney, W.J., Stevens, D.K., Doucette, W.J. (1990). Decision support model for hazardous-waste application rates at land treatment systems. *Civil Engineering Systems* 7 (4), 219–228.

Grenzeback, L.R., Reilly, W.R., Roberts, P.O., Stowers, J.R. (1990). Urban freeway gridlock study: Decreasing the effects of large trucks on peak-period urban freeway congestion. *Transportation Research Record* 1256, 16–26.

Groothuis, P.A., Miller, G. (1997). The role of social distrust in risk-benefit analysis: A study of the siting of a hazardous waste disposal facility. *Journal of Risk and Uncertainty* 15 (3), 241–257.

Haas, T.J., Kichner, J.J. (1987). Hazardous materials in marine transportation – a practical course. *Journal of Chemical Education* 64 (1), 34–35.

Hall, R.W. (1986). The fastest path through a network with random time-dependent travel times. *Transportation Science* 20 (3), 182–188.

Hanna, S.R., Chang, J.C., Strimaitis, D.G. (1993). Hazardous gas model evaluation with field observations. *Atmospheric Environment* 27A (15), 2265–2285.

Hansen, P. (1980). Bicriterion path problems. In: Fandel, G., Gal, T. (Eds.), *Multiple Criteria Decision Making: Theory and Applications*. *Lecture Notes in Economics and Mathematical Systems*, vol. 177. Springer-Verlag, Berlin, pp. 109–127.

Harwood, D.W., Russell, E.R., Viner, J.G. (1989). Characteristics of accidents and incidents in highway transportation of hazardous materials. *Transportation Research Record* 1245, 23–33.

Harwood, D.W., Viner, J.G., Russell, E.R. (1993). Procedure for developing truck accident and release rates for hazmat routing. *Journal of Transportation Engineering – ASCE* 119 (2), 189–199.

Helander, M.E., Melachrinoudis, E. (1997). Facility location and reliable route planning in hazardous material transportation. *Transportation Science* 31 (3), 216–226.

Henley, E.J., Kumamoto, H. (1981). *Reliability Engineering and Risk Assessment*. Prentice Hall, Englewood Cliffs, NJ.

Hillsman, E.L. (1988). Estimating population at risk from release of hazardous materials. In: *Seminar on Multipurpose Land Information Systems: Application of Information Technology for Natural Resource Planning, Management, and Monitoring, Univ. of Wisconsin, October 24, 1986*. Report 32, Inst. for Environmental Studies, pp. 79–96

Hobeika, A.G., Kim, S. (1993). Databases and needs for risk assessment of hazardous materials shipments by trucks. In: Moses, L.N. (Ed.), *Transportation of Hazardous Materials*. Kluwer Academic, Boston, MA, p. 146.

Hoffman, G., Janko, J. (1990). Travel times as a basic part of the LISB guidance strategy. Paper presented at the IEEE Road Traffic Control Conference, London.

Hollister, K.A.K. (2002). A risk/cost framework for logistics policy evaluation: Hazardous waste management. *Journal of Business & Economic Studies* 8 (1), 51–65.

Horlick-Jones, T. (1995). Modern disasters as outrage and betrayal. *International Journal of Mass Emergencies and Disasters* 13 (3), 305–315.

Horman, R.L. (1987). Time/loss analysis. Technical report, EG and G Idaho, Inc., Idaho Falls, ID (United States), System Safety Development Center.

Huang, B., Cheu, R.L. (2004). GIS and genetic algorithms for HAZMAT route planning with security considerations. *International Journal of Geographical Information Science* 18 (8), 769–787.

Huang, B., Long, C.R., Liew, Y.S. (2003). GIS – AHP model for HAZMAT routing with security considerations. In: *IEEE 6th Int'l Conf. on ITS* (ITSC2003).

Hwang, S.T., Brown, D.F., O'Steen, J.K., Policastro, A.J., Dunn, W.E. (2001). Risk assessment for national transportation of selected hazardous materials. *Transportation Research Record* 1763, 114–124.

Iakovou, E.T. (2001). An interactive multiobjective model for the strategic maritime transportation of petroleum products: Risk analysis and routing. *Safety Science* 39 (1–2), 19–29.

Iakovou, E., Douligeris, C., Li, H., Ip, C., Yudhbir, L. (1999). A maritime global route planning model for hazardous materials transportation. *Transportation Science* 33 (1), 34–48.

ICF Consulting (2000). Risk management framework for hazardous materials transportation. Submitted to Research and Special Programs Administration, US Department of Transportation.

Jacobs, T.L., Warmerdam, J.M. (1994). Simultaneous routing and siting for hazardous-waste operations. *Journal of Urban Planning and Development – ASCE* 120 (3), 115–131.

Jin, H.H., Batta, R. (1997). Objectives derived from viewing hazmat shipments as a sequence of independent Bernoulli trials. *Transportation Science* 31 (3), 252–261.

Jin, H.H., Batta, R.J., Karwan, M.H. (1996). On the analysis of two new models for transporting hazardous materials. *Operations Research* 44 (5), 710–723.

Jonkman, S.N., van Gelder, P.H.A.J.M, Vrijling, J.K. (2003). An overview of quantitative risk measures for loss of life and economic damage. *Journal of Hazardous Materials A* 99, 1–30.

Kalelkar, A.S., Brinks, R.E. (1978). Use of multidimensional utility functions in hazardous shipment decisions. *Accident Analysis and Prevention* 10, 251–265.

Kara, B.Y., Verter, V. (2004). Designing a road network for hazardous materials transportation. *Transportation Science* 38 (2), 188–196.

Kara, B.Y., Erkut, E., Verter, V. (2003). Accurate calculation of hazardous materials transport risks. *Operations Research Letters* 31 (4), 285–292.
Karkazis, J., Boffey, T.B. (1995). Optimal location of routes for vehicles transporting hazardous materials. *European Journal of Operational Research* 86 (2), 201–215.
Kasperson, R.E., Renn, O., Slovic, P., Brown, H.S., Emel, J., Goble, R., Kasperson, J.S., Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis* 8 (2), 177–187.
Keeney, R. (1980). Equity and public risk. *Operations Research* 28, 527–534.
Keeney, R.L., Winkler, L.R. (1985). Evaluating decision strategies for equity of public risks. *Operations Research* 33 (5), 955–970.
Kimberly, A., Killmer, H. (2002). A risk/cost framework for logistics policy evaluation: Hazardous waste management. *Journal of Business & Economic Studies* 8 (1), 51–65.
Kloeber, G., Cornell, M., McNamara, T., Moscati, A. (1979). Risk assessment of air versus other transportation modes for explosives and flammable cryogenic liquids. Volume I: Risk assessment method and results. Report DOT/RSPA/MTB-79/13, US Department of Transportation.
Koutsopoulos, H.N., Xu, H. (1993). An information discounting routing strategy for advanced traveler information systems. *Transportation Research Part C* 1 (3), 249–264.
Kulkarni, V.G. (1986). Shortest paths in networks with exponentially distributed arc lengths. *Networks* 16, 255–274.
Kunreuther, H., Easterling, D. (1990). Are risk-benefit tradeoffs possible in siting hazardous facilities. *American Economic Review* 80 (2), 252–256.
Kunreuther, H., Linnerooth, J., Vaupel, J.W. (1984). A decision-process perspective on risk and policy analysis. *Management Science* 30 (4), 475–485.
LaFrance-Linden, D., Watson, S., Haines, M.J. (2001). Threat assessment of hazardous materials transportation in aircraft cargo compartments. *Transportation Research Record* 1763, 130–137.
Larson, K.M. (1996). Hazmat on the rails: A closer look. *Environmental Solutions* 9 (9).
Lassarre, S., Fedra, K., Weigkricht, E. (1993). Computer-assisted routing of dangerous goods for Haute-Normandie. *Journal of Transportation Engineering* 119 (2), 200–210.
Leeming, D.G., Saccomanno, F.F. (1994). Use of quantified risk assessment in evaluating the risks of transporting chlorine by road and rail. *Transportation Research Record* 1430, 27–35.
Leonelli, P., Bonvicini, S., Spadoni, G. (2000). Hazardous materials transportation: A risk-analysis-Based routing methodology. *Journal of Hazardous Materials* 71, 283–300.
Lepofsky, M., Abkowitz, M., Cheng, P. (1993). Transportation hazard analysis in integrated GIS environment. *Journal of Transportation Engineering – ASCE* 119 (2), 239–254.
Levy, H. (1992). Stochastic dominance and expected utility: Survey and analysis. *Management Science* 38 (4), 555–593.
Li, H., Iakovou, E., Douliger, C. (1996). A strategic planning model for marine oil transportation in the Gulf of Mexico. *Transportation Research Record* 1522, 108–115.
Lindner-Dutton, L., Batta, R., Karwan, M. (1991). Equitable sequencing of a given set of hazardous materials shipments. *Transportation Science* 25, 124–137.
List, G., Abkowitz, M. (1986). Estimates of current hazardous material flow patterns. *Transportation Quarterly* 40, 483–502.
List, G., Mirchandani, P. (1991). An integrated network planar multiobjective model for routing and siting for hazardous materials and wastes. *Transportation Science* 25 (2), 146–156.
List, G.F., Turnquist, M.A. (1998). Routing and emergency response team siting for high-level radioactive waste shipments. *IEEE Transactions on Engineering Management* 45 (2), 141–152.
List, G.F., Mirchandani, P.B., Turnquist, M.A., Zografos, K.G. (1991). Modeling and analysis for hazardous materials transportation-risk analysis, routing scheduling and facility location. *Transportation Science* 25 (2), 100–114.
Lovett, A.A., Parfitt, J.P., Brainard, J.S. (1997). Using GIS in risk analysis: A case study of hazardous waste transport. *Risk Analysis* 17 (5), 625–633.
Luedtke, J., White, C.C. (2002). Hazmat transportation and security: Survey and directions for future research. Department of Ind. & Sys. Engineering, Georgia Institute of Technology.
Macgregor, D., Slovic, P., Mason, R.G., Detweiler, J., Binney, S.E., Dodd, B. (1994). Perceived risks of radioactive-waste transport through Oregon: Results of a statewide survey. *Risk Analysis* 14 (1), 5–14.

Madsen, O.B.G. (1983). Methods for solving combined two level location–routing problems of realistic dimensions. *European Journal of Operational Research* 12 (3), 295–301.
Magnanti, T.L., Wong, R.T. (1984). Network design and transportation planning: models and algorithms. *Transportation Science* 18, 1–55.
Marianov, V., ReVelle, C. (1998). Linear, non-approximated models for optimal routing in hazardous environments. *Journal of the Operational Research Society* 49 (2), 157–164.
Marianov, V., ReVelle, C., Shih, S. (2002). Anticoverage models for obnoxious material transportation. *Environment and Planning B – Planning & Design* 29 (1), 141–150.
Markowitz, H.M. (1987). *Mean – Variance Analysis in Portfolio Choice and Capital Markets*. Blackwell Sci., Oxford.
Marsh, M.T., Schilling, D.A. (1994). Equity measures in facility location analysis: A review and framework. *European Journal of Operations Research* 74 (1), 1–17.
Martinez-Alegria, R., Ordonez, C., Taboada, J. (2003). A conceptual model for analyzing the risks involved in the transportation of hazardous goods: Implementation in a geographic information system. *Human And Ecological Risk Assessment* 9 (3), 857–873.
McClure, T.A., Brentlinger, L.A., Drago, V.J., Kerr, D.C. (1988). Considerations in rail routing of radioactive materials. Technical report, Office of Transport Systems and Planning, Battelle Memorial Institute, Columbus, OH.
McCord, M.R., Leu, A.Y.C. (1995). Sensitivity of optimal hazmat routes to limited preference specification. *INFOR* 33 (2), 68–83.
McNeil, S., Oh, S.C. (1991). A note on the influence of rail defects on the risk associated with shipping hazardous materials by rail. *Risk Analysis* 11 (2), 333–338.
Miaou, S.P., Chin, S.M. (1991). Computing k-shortest path for nuclear spent fuel highway transportation. *European Journal of Operational Research* 53 (1), 64–80.
Milazzo, M.F., Lisi, R., Maschio, G., Antonioni, G., Bonvicini, S., Spadoni, G. (2002). HazMat transport through, Messina town: From risk analysis suggestions for improving territorial safety. *Journal of Loss Prevention in the Process Industries* 15 (5), 347–356.
Mileti, D., O'Brien, P. (1992). Warning during disasters: Normalizing communicated risk. *Social Problems* 39, 40–44.
Miller-Hooks, E.D. (2001). Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. *Networks* 37 (1), 35–52.
Miller-Hooks, E.D., Mahmassani, H.S. (1998). Optimal routing of hazardous materials in stochastic, time-varying transportation networks. *Transportation Research Record* 1645, 143–151.
Miller-Hooks, E.D., Mahmassani, H.S. (2000). Least expected time paths in stochastic, time-varying transportation networks. *Transportation Science* 34, 198–215.
Mills, G.S., Neuhauser, K.S. (1998). Urban risks of truck transport of radioactive material. *Risk Analysis* 18 (6), 781–785.
Min, H., Jayaraman, V., Srivastava, R. (1998). Combined location–routing problems: A synthesis and future research directions. *European Journal of Operational Research* 108 (1), 1–15.
Mirchandani, P.B. (1976). Shortest distance and reliability of probabilistic networks. *Computers & Operations Research* 3, 347–355.
Moore, J.A. (1989). Speaking of the data: The alar controversy. *EPA Journal* 15, 5–9.
Moore, J.E., Sandquist, G.M., Slaughter, D.M. (1995). A route-specific system for risk assessment of radioactive materials transportation accidents. *Nuclear Technology* 112 (1), 63–78.
Morales, J.M. (1989). Analytical procedures for estimating freeway traffic congestion. *TRB Research Circular* 344, 38–46.
Morin, T.L., Marsten, R.E. (1976). Branch-and-bound strategies for dynamic programming. *Operations Research* 24 (4), 611–627.
Mumpower, J.L. (1986). An analysis of the de minimis strategy for risk management. *Risk Analysis* 6, 437–446.
Nardini, L., Aparicio, L., Bandoni, A., Tonelli, S.M. (2003). Regional risk associated with the transport of hazardous materials. *Latin American Applied Research* 33 (3), 213–218.
Neill, H.R., Neill, R.H. (2000). Transportation of transuranic nuclear waste to WIPP: A reconsideration of truck versus rail for two sites. *Natural Resources Journal* 40 (1), 93–124.

Nembhard, D.A., White, C.C. (1997). Applications of non-order-preserving path selection to hazmat routing. *Transportation Science* 31 (3), 262–271.
Nguyen, S., Pallottino, S. (1986). Hyperpaths and shortest hyperpaths. In: *Combinatorial Optimization*. *Lecture Notes in Mathematics*, vol. 1403. Springer-Verlag, Berlin, pp. 258–271.
NHTSA (1996). The economic cost of motor vehicle crashes. National Highway Traffic Safety Administration, US Department of Transportation.
NIOSH (1994). NTIS publication PB-94-195047. Available at *http://www.cdc.gov/niosh/idlh/idlh-1.html*.
Nozick, L.K., List, G.F., Turnquist, M.A. (1997). Integrated routing and scheduling in hazardous materials transportation. *Transportation Science* 31 (3), 200–215.
NSC (2003). Estimating the costs of unintentional injuries, 2003. Available at *http://www.nsc.org/lrs/statinfo/estcost.htm*.
Ogryczak, W., Ruszczynski, A. (2002). Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization* 13 (1), 60–78.
Orr, M.F., Kaye, W.E., Zeitz, P., Powers, M.E., Rosenthal, L. (2001). Public health risks of railroad hazardous substance emergency events. *Journal of Occupational and Environmental Medicine* 43 (2), 94–100.
Pasquill, F., Smith, F.B. (1983). *Atmospheric Diffusion*, 3rd edition. *Ellis Horwood Series in Environmental Science*. Prentice Hall, New York.
Paté-Cornell, E. (2002). Probabilistic modeling of terrorist threats: A system analysis approach to setting priorities among counter-measures. *Military Operations Research* 7 (4), 5–23.
Patel, M.H., Horowitz, A.J. (1994). Optimal routing of hazardous materials considering risk of spill. *Transportation Research Part A – Policy and Practice* 28 (2), 119–132.
Peirce, J.J., Davidson, G.M. (1982). Linear programming in hazardous waste management. *Journal Environmental Engineering* 108 (5), 1014–1026.
Pet-Armacost, J.J., Sepulveda, J., Sakude, M. (1999). Monte Carlo sensitivity analysis of unknown parameters in hazardous materials transportation risk assessment. *Risk Analysis* 19 (6), 1173–1184.
Philipson, L.L., Napadensky, H.S., Maxey, M.N. (1983). Hazardous materials transportation risk assessment. *Transportation of Hazardous Materials: Toward a National Strategy*, vol. 2, TRB Special Report 197, pp. 43–57.
Pijawka, K.D., Foote, S., Soesilo, A., Scanlon, R.D., Cantilli, E.J. (1985). Improving transportation of hazardous materials through risk assessment and routing. Technical Report PB-86-112471/XAB; TRB/TRR-1020, National Research Council, Washington, DC (USA), Transportation Research Board.
Pine, J.C., Marx, B.D. (1997). Utilizing state hazardous materials transportation data in hazardous analysis. *Journal of Hazardous Materials* 54 (1–2), 113–122.
Puliafito, E., Guevara, M., Puliafito, C. (2003). Characterization of urban air quality using GIS as a management system. *Environmental Pollution* 122, 105–117.
Purdy, G. (1993). Risk analysis of the transportation of dangerous goods by road and rail. *Journal of Hazardous Materials* 33 (2), 229–259.
Raj, P.K., Pritchard, E.W. (2000). Hazardous materials transportation on US railroads: Application of risk analysis methods to decision making in development of regulations. *Transportation Research Record* 1707, 22–26.
Renn, O., Burns, W.J., Kasperson, J.X., Kasperson, R.E., Slovic, P. (1992). The social amplification of risk: Theoretical foundations and empirical applications. *Journal of Social Issues* 48, 137–160.
ReVelle, C., Cohon, J., Shobrys, D. (1991). Simultaneous siting and routing in the disposal of hazardous wastes. *Transportation Science* 25 (2), 138–145.
Rhyne, W.R. (1994). *Hazardous Materials Transportation Risk Analysis*. Van Nostrand–Reinhold, New York.
Roeleven, D., Kok, M., Stipdonk, H.L., Devries, W.A. (1995). Inland waterway transport – modeling the probability of accidents. *Safety Science* 19 (2–3), 191–202.
Romer, H., Haastrup, P., Petersen, H.J.S. (1995). Accidents during marine transport of dangerous goods – distribution of fatalities. *Journal of Loss Prevention in the Process Industries* 8 (1), 29–34.
Rosmuller, N., Van Gelder, P.H.A.J.M. (2003). Hazardous materials release analysis: Probabilistic input for emergency response organizations. In: Bedford, T., Van Gelder, P.H.A.J.M. (Eds.), *Safety and Reliability*. A.A. Balkema, pp. 1337–1344.

Saccomanno, F., Chan, A. (1985). Economic evaluation of routing strategies for hazardous road shipments. *Transportation Research Record* 1020, 12–18.

Saccomanno, F.F., El-Hage, S. (1989). Minimizing derailments of railcars carrying dangerous commodities through effective marshaling strategies. *Transportation Research Record* 1245, 34–51.

Saccomanno, F.F., Elhage, S.M. (1991). Establishing derailment profiles by position for corridor shipments of dangerous goods. *Canadian Journal of Civil Engineering* 18 (1), 67–75.

Saccomanno, F., Haastrup, P. (2002). Influence of safety measures on the risks of transporting dangerous goods through road tunnels. *Risk Analysis* 22 (6), 1059–1069.

Saccomanno, F.F., Shortreed, J.H. (1993). Hazmat transport risks – societal and individual perspectives. *Journal of Transportation Engineering – ASCE* 119 (2), 177–188.

Saccomanno, F.F., Shortreed, J.H., Van Aerde, M., Higgs, J. (1989). Comparison of risk measures for the transport of dangerous commodities by truck and rail. *Transportation Research Record* 1245, 1–13.

Sandquist, G.M., Bennion, J.S., Moore, J.E., Slaughter, D.M. (1993). A route-specific transportation risk assessment model. *Transactions of the American Nuclear Society* 68 (9), 151–153.

Scanlon, R.D., Cantilli, E.J. (1985). Assessing the risk and safety in the transportation of hazardous materials. *Transportation Research Record* 1020.

Seguin, R., Potvin, J.-Y., Gendreau, M., Crainic, T.G., Marcotte, P. (1997). Real-time decision problems: An operational research perspective. *Journal of the Operational Research Society* 48 (2), 162–174.

Shappert, L.B., Brobst, W.A., Langhaar, J.W., Sisler, J.A. (1973). Probability and consequences of transportation accidents involving radioactive-material shipments in the nuclear fuel cycle. *Nuclear Safety* 14 (6), 597–604.

Sherali, H.D., Brizendine, L.D., Glickman, T.S., Subramanian, S. (1997). Low-probability high-consequence considerations in routing hazardous materials shipments. *Transportation Science* 31, 237–251.

Shobrys, D. (1981). A model for the selection of shipping routes and storage locations for a hazardous substance. PhD thesis, Johns Hopkins University, Baltimore.

Sissell, K. (1995). Defining risk assessment. *Chemical Week* (September 27), S16.

Sivakumar, R.A., Batta, R. (1994). The variance-constrained shortest path problem. *Transportation Science* 28 (4), 309–316.

Sivakumar, R.A., Batta, R., Karwan, M.H. (1993). A network-based model for transporting extremely hazardous materials. *Operations Research Letters* 13 (2), 85–93.

Sivakumar, R.A., Batta, R., Karwan, M.H. (1995). A multiple route conditional risk model for transporting hazardous materials. *INFOR* 33 (1), 20–33.

Slovic, P., Lichtenstein, S., Fischhoff, B. (1984). Modeling the societal impact of fatal accidents. *Management Science* 30, 464–474.

Smith, R.D. (1987). Routing and scheduling of radioactive material shipments. PhD thesis, Texas University, Austin, USA.

Spadoni, G., Leonelli, P., Verlicchi, P., Fiore, R. (1995). A numerical procedure for assessing risks from road transport of dangerous substances. *Journal of Loss Prevention in the Process Industries* 8 (4), 245–252.

Stowers, C.L., Palekar, U.S. (1993). Location models with routing considerations for a single obnoxious facility. *Transportation Science* 27 (4), 350–362.

Sulijoadikusumo, G.S., Nozick, L.K. (1998). Multiobjective routing and scheduling of hazardous materials shipments. *Transportation Research Record* 1613, 96–104.

Swoveland, C. (1987). Risk analysis of regulatory options for the transport of dangerous commodities by rail. *Interfaces* 17 (4), 90–107.

Tayi, G.K., Rosenkrantz, D.J., Ravi, S.S. (1999). Path problems in networks with vector-valued edge weights. *Networks* 34 (1), 19–35.

Transport Canada (2004). On the move – keeping Canadians safe. Available at *http://www.tc.gc.ca/Publications/TP14217e/onthemove-e.htm*.

TRB (1993). Hazardous materials shipment information for emergency response. Special Report 239, Transportation Research Board.

TRB (2002). Deterrence, protection, and preparation: The new transportation security imperative. Special Report 270, Transportation Research Board.

Turnquist, M. (1993). Multiple objectives, uncertainty and routing decisions for hazardous materials shipments. In: Cohn, L.F. (Ed.), *Computing in Civil and Building Engineering: Proceedings of the 5th International Conference on Computing in Civil and Building Engineering*. ASCE, New York, pp. 357–364.

UN (2001). UN recommendation on the transport of dangerous goods, model regulations. United Nations Economic and Social Council's Committee of Experts on the Transport of Dangerous Goods.

Urbanek, G.I., Barber, E.J. (1980). Development of criteria to designate routes for transporting hazardous materials. Final Report FHWA–RD–80-105, Federal Highway Administration, Washington, DC.

US DOT (1994). Guidelines for applying criteria to designate routes for transporting hazardous materials. Report FHWA-SA-94-083, US Department of Transportation, Federal Highway Administration, Washington, DC.

US DOT (1998). Hazardous materials shipments. Office of Hazardous Materials Safety, Research and Special Programs Administration, US Department of Transportation, Washington, DC.

US DOT (2000). Department wide program evaluation of the hazardous materials transportation programs. The Office of Hazardous Materials Safety, US Department of Transportation, Washington, DC.

US DOT (2004a). 2002 and 2003 summary of hazardous materials transportation incidents. Research and Special Programs Administration, Office of Hazardous Materials Safety, US Department of Transportation, Washington, DC. Available at *http://hazmat.dot.gov/files/summary/2003/2003sum.pdf*.

US DOT (2004b). List of hazardous materials. The Office of Hazardous Materials Safety, US Department of Transportation, Washington, DC.

US DOT (2004c). Hazmat summary by mode/cause: Calendar year 2003. Serious incidents, printed on January 4, 2005. The Office of Hazardous Materials Safety, US Department of Transportation, Washington, DC. Available at *http://hazmat.dot.gov/pubs/inc/data/2003/2003scause.pdf*.

US EPA (2004). Computer-aided management of emergency operations software with ALOHA. US Environmental Protection Agency. Complete Kit available at *http://response.restoration.noaa.gov/cameo/aloha.html*.

Vanaerde, M., Shortreed, J., Stewart, A.M., Matthews, M. (1989). Assessing the risks associated with the transport of dangerous goods by truck and rail using the riskmod model. *Canadian Journal of Civil Engineering* 16 (3), 326–334.

Vansteen, J.F.J. (1987). A methodology for aiding hazardous materials transportation decisions. *European Journal of Operational Research* 32 (2), 231–244.

Verma, M., Verter, V. (2007). Rail transportation of hazardous materials: Population exposure to airborne toxins. *Computers & Operations Research* 34 (5), 1287–1303.

Verter, V., Erkut, E. (1995). Hazardous materials logistics: An annotated bibliography. In: Haurie, A., Carraro, C. (Eds.), *Operations Research and Environmental Management*. Kluwer Academic, pp. 221–267.

Verter, V., Erkut, E. (1997). Incorporating insurance costs in hazardous materials routing models. *Transportation Science* 31 (3), 227–236.

Verter, V., Kara, B.Y. (2001). A GIS-based framework for hazardous materials transport risk assessment. *Risk Analysis* 21 (6), 1109–1120.

Verter, V., Kara, B.Y. (2005). A path-based approach for the hazardous network design problem. Working paper, Faculty of Management, McGill University.

Vesely, W.E., Goldberg, F.F., Roberts, N.H., Haasl, D.F. (1981). Fault tree handbook. NUREG - 0492. US Nuclear Regulatory Commission.

Weigkricht, E., Fedra, K. (1995). Decision-support systems for dangerous goods transportation. *INFOR* 33 (2), 84–99.

Wijeratne, A.B., Turnquist, M.A., Mirchandani, P.B. (1993). Multiobjective routing of hazardous materials in stochastic networks. *European Journal of Operational Research* 65, 33–43.

Wilmot, E.L., Cashwell, J.W., Joy, D.S. (1983). Analysis of population densities along transportation routes. *The 7th international symposium on packing and transportation of radioactive materials*, New Orleans, LA, USA, p. 10.

Wirasinghe, S. (1978). Determination of traffic delays from shock-wave analysis. *Transportation Research* 12, 343–348.

Yang, B. (2001). Robust on-line routing in intelligent transportation systems. PhD dissertation, Department of Civil and Environmental Engineering, The Pennsylvania State University.

Yitzhaki, S. (1982). Stochastic dominance, mean variance, and Gini's mean difference. *The American Economic Review* 72, 178–185.

Zhang, J.J., Hodgson, J., Erkut, E. (2000). Using GIS to assess the risks of hazardous materials transport in networks. *European Journal of Operational Research* 121 (2), 316–329.

Zografos, K.G., Androutsopoulos, K.N. (2001). Assessing impacts from introduction of advanced transport telematics technologies in hazardous materials fleet management. In: *Proceedings of the 80th Annual Meeting of Transportation Research Board*, Washington DC, USA.

Zografos, K.G., Androutsopoulos, K.N. (2002). Heuristic algorithms for solving hazardous materials logistical problems. *Transportation Research Record* 1783, 158–166.

Zografos, K.G., Androutsopoulos, K.N. (2004). A heuristic algorithm for solving hazardous materials distribution problems. *European Journal of Operational Research* 152 (2), 507–519.

Zografos, K.G., Davis, C. (1989). Multi-objective programming approach for routing hazardous materials. *Journal of Transportation Engineering* 115, 661–673.

Zografos, K.G., Samara, S. (1989). Combined location–routing model for hazardous waste transportation and disposal. *Transportation Research Record* 1245, 52–59.

Zografos, K.G., Vasilakis, G.M., Giannouli, I.M. (2000). A methodological framework for developing a DSS for hazardous material emergency response operations. *Journal of Hazardous Materials* 71, 503–521.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14010-4

Chapter 10
Traffic Equilibrium

Patrice Marcotte

Department of Computer Science and Operations Research, University of Montreal, Montreal, QC, Canada H3C 3J7
E-mail: marcotte@iro.umontreal.ca

Michael Patriksson

Department of Mathematics, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden
E-mail: mipat@math.chalmers.se

1 Background

The subject of traffic equilibrium is the description, through analytical tools, of the stationary distribution of vehicles in a transportation network. Assuming that travelers seek to minimize their individual travel cost, an equilibrium is reached when no traveler has an incentive to modify its travel decision. Historically, the term *traffic assignment* was used to describe the same phenomenon, reflecting the fact that the practice was not so much of estimating the traffic distribution through analytical models – including design or pricing aspects – than performing an assignment of travelers onto the network, typically in order to assess the performance of traffic control policies.

Traffic equilibrium is the cornerstone, or the 'inner loop', of any modern network analysis. Its definition dates back to the 1950s and hundreds of papers have been devoted to its study. In the last two decades, it has served as the quintessential benchmark by which the performance of equilibrium algorithms could be measured, and became a field of research of its own. In fact, much of the early development of algorithms for finite-dimensional variational inequality problems defined over convex sets took place within this field, and at one point or another several of the foremost researchers in mathematical programming and operations research have contributed to the theory of traffic equilibrium. Presenting a comprehensive account of the traffic equilibrium problem (or TEP, for short) would however require an entire book and, since such books already exist, is beyond the scope of the current chapter.

In the 21st century, we understand rather well how to solve the basic TEP; consequently, the challenges and interests of the transportation community have shifted toward enhanced models that generalize the basic equilibrium model. In this chapter, the authors briefly present the main theoretical and algorithmical results pertaining to the TEP, along the way improving theoretical results that were established some 20 years ago, and then focus their attention

on topics that have been overlooked, such as the relationship between the Nash and Wardrop concepts, as well as on new paradigms built around the basic TEP. In that respect, the presentation style and choice of topics are highly personal, and strongly reflect the authors' inclinations. Each section is completed by a 'Bibliographical notes' section, where we outline the relevant literature and briefly mention topics not covered in the main text. Finally, the appendices provide a list of notation, as well as a primer on variational inequalities, which constitute the adequate framework for modeling network equilibrium problems, as was recognized independently by Stella Dafermos and Michael J. Smith more than 25 years ago.

1.1 Bibliographical notes

Most books on quantitative transportation analysis discuss the issue of traffic assignment. A book entirely devoted to the topic has been written by Patriksson (1994b). Among several surveys in handbooks, let us mention that of Florian and Hearn (1995). These works all assume that costs increase with volumes, although this does not hold in reality, where flow rates may *decrease* as the network becomes congested. Although such effects are difficult to assess in a static model, let us mention the proposal by Nesterov (2000) for addressing this issue in situations where congestion levels are not critical.

Much of the analysis in this chapter can be directly transposed from traffic networks to computer communication networks, where origin–destination pairs are connected computer communication centers, and where vehicles correspond to message packets being routed between them. In the fixed demand case – the most natural framework in computer communication networks – the goal is to minimize total (that is, average) delay, which usually is modeled as that of minimizing the function

$$\phi(\mathbf{v}) := \sum_{l \in \mathcal{L}} \frac{v_l}{c_l - v_l} + p_l v_l,$$

where v_l denotes the arrival rate of packets at link l, c_l is the transmission capacity of the link, and p_l is the processing and propagation delay on the link. Provided that the capacities c_l are sufficiently large, i.e., capacity constraints are not tight at equilibrium, most of the results presented in this chapter go through with few alterations.

2 The basic theme

This section is concerned with the basic traffic equilibrium model, including the elastic demand case, together with its many mathematical statements. Indeed, Wardrop's conditions, which express a variational principle, can be formulated in terms of a complementarity, variational inequality or optimization program, either in terms of route or link flows. Throughout the section, key

notations and definitions are introduced, while a small numerical example will help in absorbing them. A comprehensive list of the notation is provided in Appendix B.

2.1 The Wardrop conditions

The basic elements of the standard traffic equilibrium model are: (i) a transportation network, (ii) travel requirements, and (iii) cost functions,[1] from which traffic volumes (or flows), expressed as vehicular rates, must be deduced. More precisely, let us consider a directed graph $\mathcal{G} := (\mathcal{N}, \mathcal{L})$, consisting of a set of nodes $i \in \mathcal{N}$ and directed links $l \in \mathcal{L}$, sometimes also denoted $l = (i, j)$ by the tail and head nodes i and j, respectively. Consider also a set $\mathcal{C} \subset \mathcal{N} \times \mathcal{N}$ of origin–destination (OD) pairs (p, q), defining starting and ending nodes of network trips in $\mathcal{G}$. Throughout the section we consider an example based on the graph illustrated in Figure 1, that involves five nodes, eight links and two OD pairs: the first has node 1 as starting node and node 5 as ending node, and the second OD pair has starting and ending nodes 2 and 4, respectively.

Any well-founded traffic model recognizes the individual network user's right to decide when, where, and how to travel. A traffic equilibrium model, in which one aims at providing a macroscopic description or prediction of the traffic volume resulting from route choices made in the traffic network, must therefore be based on a sound route-choice behavioral principle.

The *equilibrium condition* refers to the concept initially introduced by the statistician J.G. Wardrop of the British Road Research Laboratory. Ever since his seminal paper, user equilibrium conditions have been known as *Wardrop's first principle*. The precise definition is as follows:

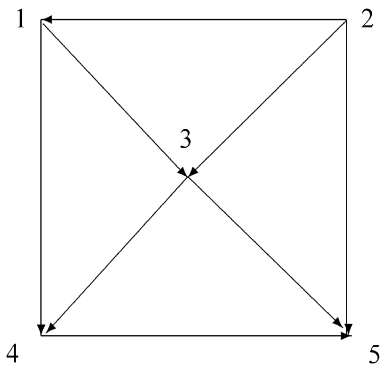


Fig. 1. An example traffic network.

[1] Throughout the chapter, the term 'cost' refers to the disutility experienced by a user of the network. In the basic models, it is a synonym for 'travel time' or 'travel delay'.

> The journey times on all the routes actually used are equal, and not greater than those which would be experienced by a single vehicle on any unused route.

Wardrop also introduced a second principle, whereby users choose their routes such that the *average* (or *total*) travel cost is minimal. This concept will be used in Section 2.9.

To translate the concept of user equilibrium into mathematical terms, we denote by $\mathcal{R}_{pq}$ the (finite) set of simple (cycle-free) routes for OD pair (p, q), by h_r the volume of traffic on route $r \in \mathcal{R}_{pq}$, and by c_r the travel cost on the route as experienced by an individual user, given the current volume of traffic. Fixing the travel costs to these values, the user equilibrium conditions can be written as follows:

$$h_r > 0 \quad \Longrightarrow \quad c_r = \pi_{pq}, \quad r \in \mathcal{R}_{pq}, \ p, q \in \mathcal{C}, \tag{1a}$$

$$h_r = 0 \quad \Longrightarrow \quad c_r \geqslant \pi_{pq}, \quad r \in \mathcal{R}_{pq}, \ p, q \in \mathcal{C}, \tag{1b}$$

where π_{pq} denotes the minimal (that is, equilibrium) route cost for OD pair (p, q).

These conditions express the optimality conditions of a shortest route problem for each OD pair, where route costs are given by c_r. What makes the problem more complex than a shortest route problem is that the supply and demand characteristics are not fixed: normally, the route costs c_r depend on the volume of traffic on the routes, and the total traffic volume in an OD pair (that is, the OD demand) may depend on the least cost π_{pq} of travel, perhaps even upon other OD pairs' costs as well. In order for the above system to describe an equilibrium state, we must further incorporate the cost perception of the users given the volume of traffic, and the mechanism by which flow demand is generated.

Let therefore $c_r : \Re_+^{|\mathcal{R}|} \to \Re$ be a real-valued function, describing the cost of utilizing route $r \in \mathcal{R}$, and whose argument is the vector $\mathbf{h} \in \Re^{|\mathcal{R}|}$ of route flows h_r. The vector of these route costs forms the aggregated vector-valued function $\mathbf{c} : \Re_+^{|\mathcal{R}|} \to \Re^{|\mathcal{R}|}$. Further, we assume that the demand on OD pair $(p, q) \in \mathcal{C}$ is given by the real-valued, nonnegative function $g_{pq} : \Re^{|\mathcal{C}|} \to \Re_+$, whose argument is the vector $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$ of OD pair least travel costs, π_{pq}, introduced in (1). The vector-valued demand function of the least travel costs is denoted by $\mathbf{g} : \Re^{|\mathcal{C}|} \to \Re_+^{|\mathcal{C}|}$. We also introduce the route–OD pair incidence matrix $\boldsymbol{\Gamma} \in \Re^{|\mathcal{R}| \times |\mathcal{C}|}$ whose element γ_{rk} is set to 1 if route r joins OD pair $k = (p, q) \in \mathcal{C}$, and 0 otherwise.

Based on the above notation, one may express the conditions (1) and the demand constraints as the system:

$$\mathbf{0}^{|\mathcal{R}|} \leqslant \mathbf{h} \perp \left(\mathbf{c}(\mathbf{h}) - \boldsymbol{\Gamma}\boldsymbol{\pi}\right) \geqslant \mathbf{0}^{|\mathcal{R}|}, \tag{2a}$$

$$\boldsymbol{\Gamma}^\top \mathbf{h} = \mathbf{d}, \tag{2b}$$

$$\mathbf{d} = \mathbf{g}(\boldsymbol{\pi}), \tag{2c}$$

where, for two arbitrary vectors $\mathbf{a}, \mathbf{b} \in \Re^n$, the notation $\mathbf{a} \perp \mathbf{b}$ means that $\mathbf{a}^\top \mathbf{b} = 0$. An interesting special case of the above conditions occurs when demand is *fixed*, that is, when $\mathbf{g} \equiv \bar{\mathbf{d}} \in \Re_+^{|\mathcal{C}|}$. In the sequel, we denote by H_d the set of couples $(\mathbf{h}, \mathbf{d})$ that satisfy equalities (2b)–(2c), and by H the corresponding set when demand is fixed.

Formulation (2) is a special instance of a *mixed complementarity* problem; it is the basic traffic equilibrium model from which we will derive several equivalent ones, in the forms of variational inequality, nonlinear complementarity, and optimization problems. This framework involves the pair $(\mathbf{h}, \boldsymbol{\pi})$ and the demand vector $\mathbf{d} \in \Re^{|\mathcal{C}|}$, although one could substitute its value $\mathbf{g}(\boldsymbol{\pi})$ to the latter. If demand is fixed, condition (2c) vanishes.

The notion of equilibrium, as described in (2), should be thought of as a steady-state evolving after a transient (disequilibrium) phase in which travelers successively adjust their route choices, seeking to minimize travel costs under prevailing traffic conditions, until a situation with stable route travel costs and route flows has been reached.

2.2 Link flow representations

As the number of routes in a practical application can be enormous, we are interested in developing representations of the Wardrop conditions that are defined in terms of link flows only.

Suppose that for every route $r \in \mathcal{R}$, the route cost $c_r(\mathbf{h})$ is *additive*, i.e., $c_r(\mathbf{h})$ is the sum of the costs of using all links defining r. Let $\boldsymbol{\Lambda} \in \{0, 1\}^{|\mathcal{L}| \times |\mathcal{R}|}$ be the link–route incidence matrix, whose element λ_{lr} equals one if route $r \in \mathcal{R}$ utilizes link $l \in \mathcal{L}$, and zero otherwise. We can then write $c_r(\mathbf{h}) = \sum_{l \in \mathcal{L}} \lambda_{lr} t_l(\mathbf{v})$, or, in compact form,

$$\mathbf{c}(\mathbf{h}) = \boldsymbol{\Lambda}^\top \mathbf{t}(\mathbf{v}), \tag{3}$$

where $\mathbf{v} \in \Re^{|\mathcal{L}|}$ is the vector of total volumes of traffic on the links. The cost $t_l(\mathbf{v})$ of traversing link $l \in \mathcal{L}$ at the flow $\mathbf{v}$ is given by a function $t_l : \Re_+^{|\mathcal{L}|} \to \Re$; the vector of these link costs define the function $\mathbf{t} : \Re_+^{|\mathcal{L}|} \to \Re^{|\mathcal{L}|}$.

Implicit in the relationship (3) is the assumption that the pair $(\mathbf{h}, \mathbf{v})$ is consistent, in the sense that $\mathbf{v}$ equals the sum of the route flows:

$$v_l = \sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \lambda_{lr} h_r, \quad l \in \mathcal{L},$$

or, in vector-matrix form,

$$\mathbf{v} = \boldsymbol{\Lambda} \mathbf{h}. \tag{4}$$

From the above we derive the formula $\mathbf{c}(\mathbf{h}) = \boldsymbol{\Lambda}^\top \mathbf{t}(\boldsymbol{\Lambda} \mathbf{h})$ for evaluating a set of route costs consistent with given route flows. Adding the two equations (3) and (4) to (2) yields a formulation where route flow variables are relegated to the constraints.

In our example, the link flow vector $\mathbf{v}$ is $\mathbf{v} = (v_{13}, v_{14}, v_{21}, v_{23}, v_{25}, v_{34}, v_{35}, v_{45})^\top$. The graph in Figure 1 then corresponds to the following link–route incidence matrix: in the first OD pair we identify three loop-free routes, namely the node ordering $1 \to 3 \to 5$, $1 \to 4 \to 5$, and $1 \to 3 \to 4 \to 5$, while the second OD pair has the three routes $2 \to 1 \to 4$, $2 \to 3 \to 4$, and $2 \to 1 \to 3 \to 4$. The corresponding link–route matrix $\mathbf{\Lambda} \in \Re^{8\times 6}$ and vector $\mathbf{t}$ of link cost functions are defined as

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{and}$$

$$\mathbf{t}(\mathbf{v}) = \begin{pmatrix} t_{13}(v_{13}) \\ t_{14}(v_{14}) \\ t_{21}(v_{21}) \\ t_{23}(v_{23}) \\ t_{25}(v_{25}) \\ t_{34}(v_{34}) \\ t_{35}(v_{35}) \\ t_{45}(v_{45}) \end{pmatrix} := \begin{pmatrix} 1 + \frac{1}{2}v_{13}, \\ 1 + v_{14}, \\ 1 + 5v_{21}, \\ 2 + 3v_{23}, \\ 4 + v_{25}, \\ 3 + v_{34}, \\ 2 + 2v_{35}, \\ 1 + 4v_{45} \end{pmatrix}.$$

In this network we also define the (separable) OD demand functions $g_{15}(\pi_{15}) := 11 - \pi_{15}$ and $g_{24}(\pi_{24}) := 11 - \pi_{24}$, with inverses $\pi_{15} = g^{-1}(d_{15}) =: \xi_{15}(d_{15}) = 11 - d_{15}$ and $\pi_{24} = g^{-1}(d_{24}) =: \xi_{24}(d_{24}) = 11 - d_{24}$, respectively.

The reader is invited to check that $\boldsymbol{\pi}^* = (8, 9)^\top$, with $\mathbf{g}(\boldsymbol{\pi}^*) = (3, 2)^\top$. Further, a vector of equilibrium route flows is $\mathbf{h}^* = (h^*_{11}, h^*_{12}, h^*_{13}, h^*_{21}, h^*_{22}, h^*_{23})^\top = (2, 1, 0, 1, 1, 0)^\top$, and a vector of equilibrium link flows is $\mathbf{v}^* = (2, 2, 1, 1, 0, 1, 2, 1)^\top$. The corresponding link costs are $\mathbf{t}(\mathbf{v}^*) = (2, 3, 6, 5, 4, 4, 6, 5)^\top$. Regarding the equilibrium conditions, we observe, for the first OD pair, that the route costs are 8, 8, and 11, respectively. According to the equilibrium principle, the more costly third route will not be used. Similarly, in the second OD pair, the route costs are 9, 9, and 12, respectively. The reader may check that the equations $\mathbf{v}^* = \mathbf{\Lambda}\mathbf{h}^*$ and $\mathbf{c}(\mathbf{h}^*) = \mathbf{\Lambda}^\top \mathbf{t}(\mathbf{v}^*)$ are satisfied.

Another formulation involves the origin–destination flows, also referred to as *commodity flows*,[2] and dispenses with route flow variables. Towards this aim, we introduce the *link–node incidence matrix* $\mathbf{E} \in \{-1, 0, 1\}^{|\mathcal{N}|\times|\mathcal{L}|}$, whose ele-

[2] The term *commodity* stems from the economics literature, where an OD flow is associated with a specific good.

ment e_{il} equals -1 if node i is the origin node of link l, 1 if node i is the destination node of link l, and 0 otherwise. The link–node version of Wardrop's equilibrium conditions states that at an equilibrium link flow $\mathbf{v}$, equal to the aggregate of commodity (OD pair) volumes $\mathbf{w}_k \in \Re^{|\mathcal{L}|}$, $k := (p, q) \in \mathcal{C}$, there exist vectors $\boldsymbol{\pi}_k \in \Re^{|\mathcal{N}|}$, $k \in \mathcal{C}$, of node prices (alternatively node potentials or dual variables) such that for a given link $(i, j) \in \mathcal{L}$,

$$0 \leqslant w_{ijk} \perp \big(t_{ij}(\mathbf{v}) - [\pi_{jk} - \pi_{ik}]\big) \geqslant 0, \quad k \in \mathcal{C}, \tag{5}$$

or

$$\mathbf{0}^{|\mathcal{L}|} \leqslant \mathbf{w}_k \perp \big(\mathbf{t}(\mathbf{v}) - \mathbf{E}^\top \boldsymbol{\pi}_k\big) \geqslant \mathbf{0}^{|\mathcal{L}|}, \quad k \in C. \tag{6a}$$

To confirm the agreement with (2a), we select any OD pair $k = (p, q) \in \mathcal{C}$ and a route $r \in \mathcal{R}_{pq}$, and consider a consistent set of volumes $\mathbf{w}$, $\mathbf{v}$, and $\mathbf{h}$. Summing the above conditions (5) over route r, we obtain that

$$\sum_{l=(i,j)\in\mathcal{L}} \lambda_{lr}\big(t_l(\mathbf{v}) - [\pi_{jk} - \pi_{ik}]\big) = c_r(\mathbf{h}) - [\pi_{qk} - \pi_{pk}],$$

and since $\sum_{l=(i,j)\in\mathcal{L}} \lambda_{lr} w_{kl} = h_r$ can be made to hold by the Flow Decomposition Theorem 2 (see below), we conclude that (2) and (6) are equivalent, provided we identify π_k with $\pi_{qk} - \pi_{pk}$.

Indeed, let us introduce the indicator vector $\mathbf{i}_k \in \{-1, 0, 1\}^{|\mathcal{N}|}$, which is zero in all positions but two where, by the sign convention introduced earlier for the incidence matrix $\mathbf{E}$, the element with value 1 (-1) corresponds to the sink (source) node. The demand-feasibility relation (2b) is then expressed as

$$\mathbf{E}\mathbf{w}_k = \mathbf{i}_k g_k\big(\big[\mathbf{i}_\kappa^\top \boldsymbol{\pi}_\kappa\big]_{(\kappa\in\mathcal{C})}\big), \quad k \in \mathcal{C}. \tag{6b}$$

In the case of fixed demands, the latter right-hand sides are simply replaced by $\mathbf{i}_k \bar{d}_k$, where $\bar{d}_k \in \Re_+$ is the fixed demand for the OD pair k. We will later establish under which additional conditions the representations (2) and (6) of Wardrop's user equilibrium conditions are equivalent.

Some comments are in order. First, the vector $\mathbf{i}_k$ relates the OD node price vectors $\boldsymbol{\pi}_k \in \Re^{|\mathcal{N}|}$ appearing in the link–node flow representation (6) and the OD least cost vector $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$ appearing in the link–route version (2), through the relation

$$\boldsymbol{\pi} = (\pi_{pq})_{(p,q)\in\mathcal{C}} = \big(\mathbf{i}_k^\top \boldsymbol{\pi}_k\big)_{(k\in\mathcal{C})};$$

this is the same as stating that, for each $k \in \mathcal{C}$, $\pi_k = \pi_{qk} - \pi_{pk}$, as discussed above.

This relationship allows for a simplification of the right-hand side of (6b). Further, the vectors $\mathbf{i}_k$ clarify the relationships between the respective demand vectors. In link-flow based transportation problems, the (fixed) demand in some commodity k is denoted by an $|\mathcal{N}|$-vector $\mathbf{d}_k$, wherein a positive

(negative) element represents a sink (source) node and a null element a transshipment node. In the traffic equilibrium problem, commodities are normally associated with OD pairs, whence the vector $\mathbf{d}_k$ just mentioned equals the $|\mathcal{N}|$-vector $\mathbf{i}_k \bar{d}_k$, which has precisely two nonzero entries.

Returning to our numerical example, the node–link incidence matrix $\mathbf{E} \in \Re^{5\times 8}$ is

$$\mathbf{E} = \begin{pmatrix} -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

A commodity link flow in equilibrium is given by $\mathbf{w}_1^* = (0, 1, 1, 1, 0, 1, 0, 0)^\top$ and $\mathbf{w}_2^* = (2, 1, 0, 0, 0, 0, 2, 1)^\top$; notice that $\mathbf{w}_1^* + \mathbf{w}_2^* = \mathbf{v}^*$. Further, we can relate the equilibrium OD travel costs $\boldsymbol{\pi}^* = (9, 8)^\top$ to the node prices associated with the above network at equilibrium. Under the natural ordering of the nodes, the node price vector for the respective OD pair is

$$\boldsymbol{\pi}_1^* = \begin{pmatrix} 0 \\ +\infty \\ 2 \\ 3 \\ 8 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\pi}_2^* = \begin{pmatrix} 6 \\ 0 \\ 5 \\ 9 \\ 5 \end{pmatrix},$$

as a simple calculation, (in principle) using Dijkstra's algorithm, would show. With the vectors $\mathbf{i}_1 = (-1, 0, 0, 0, 1)^\top$ and $\mathbf{i}_2 = (0, -1, 0, 1, 0)^\top$, we see that the difference in node prices between the ending and starting nodes of the two OD pairs yield precisely the elements of $\boldsymbol{\pi}^* = (9, 8)^\top$, which confirms the above connection. This also shows the familiar behavior of node prices in network flow models: their values are only given up to an arbitrary common additive constant; we have resolved this problem by setting the node price for the starting node in each OD pair to zero.

In Section 2.7 we show that suitable properties of the demand and travel cost functions imply that the entities $\boldsymbol{\pi}^*$, $\mathbf{d}^* = \mathbf{g}(\boldsymbol{\pi}^*)$, and $\mathbf{v}^*$ are unique at equilibrium, while $\mathbf{h}^*$, and $\mathbf{w}^*$ are not guaranteed to be unique. The reader is asked to verify that in our case they are unique regardless.

2.3 Variational inequality representations

Variational inequalities provide a convenient framework for the modeling of equilibrium problems, and are briefly reviewed in Appendix A. From now on we often use the short-hand 'VIP' to refer to a variational inequality problem. The notation VIP($\mathbf{f}, X$) refers to a variational inequality problem of the following form: find a vector $\mathbf{x}^* \in X$ such that

$$\mathbf{f}(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geqslant 0, \quad \mathbf{x} \in X,$$

where $X \subseteq \Re^n$ is nonempty and closed and $\mathbf{f}: \Re^n \to \Re^n$ is continuous. An equivalent statement of the variational inequality problem that will also often be used is the following:

$$-\mathbf{f}(\mathbf{x}^*) \in N_X(\mathbf{x}^*),$$

where $N_X(\mathbf{x})$ is the *normal cone* to the set X at $\mathbf{x} \in X$. See also (94).

Link-based variational inequality models are readily developed from model (2) by incorporating the definitional constraints (4), and from the model (6) by adding the compatibility constraint

$$\mathbf{v} = \sum_{k \in \mathcal{C}} \mathbf{w}_k. \tag{7}$$

Starting from the link–route representation, the fixed demand model is VIP($\mathbf{t}, \widehat{F}$), where

$$\widehat{F} := \left\{ \mathbf{v} \in \Re^{|\mathcal{L}|} \mid \exists \mathbf{h} \in H \text{ with } \mathbf{v} = \mathbf{\Lambda h} \right\}.$$

In the case of invertible demand functions, we can write the elastic model as

$$\left[-\mathbf{t}(\mathbf{v}), \boldsymbol{\xi}(\mathbf{d}) \right] \in N_{\widehat{F}_d}(\mathbf{v}, \mathbf{d}), \tag{8}$$

where

$$\widehat{F}_d := \left\{ (\mathbf{v}, \mathbf{d}) \in \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \mid \exists (\mathbf{h}, \mathbf{d}) \in H_d \text{ with } \mathbf{v} = \mathbf{\Lambda h} \right\}.$$

The corresponding link–node representations are as follows. The fixed demand model is VIP($\mathbf{t}, F$), with

$$F := \left\{ \mathbf{v} \in \Re^{|\mathcal{L}|} \;\middle|\; \exists \mathbf{w}_k \in \Re_+^{|\mathcal{L}|} \text{ with } \mathbf{E}\mathbf{w}_k = \mathbf{i}_k \bar{d}_k, k \in \mathcal{C}, \text{ and } v_l = \sum_{k \in \mathcal{C}} w_{lk}, l \in \mathcal{L} \right\}$$

and the elastic demand model is

$$\left[-\mathbf{t}(\mathbf{v}), \boldsymbol{\xi}(\mathbf{d}) \right] \in N_{F_d}(\mathbf{v}, \mathbf{d}), \tag{9}$$

where

$$F_d := \left\{ (\mathbf{v}, \mathbf{d}) \in \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \;\middle|\; \exists \mathbf{w}_k \in \Re_+^{|\mathcal{L}|} \text{ with } \mathbf{E}\mathbf{w}_k = \mathbf{i}_k d_k, k \in \mathcal{C}, \text{ and } v_l = \sum_{k \in \mathcal{C}} w_{lk}, l \in \mathcal{L} \right\}.$$

Note that both sets $\widehat{F}$ and $\widehat{F}_d$ are implicitly defined, and therefore not available in closed form (since the routes in $\mathcal{R}$ normally are not enumerated and therefore unavailable), whereas the sets F and F_d are explicit.

2.4 A fixed demand reformulation of the elastic demand model

Suppose that the demand function $\mathbf{g}$ is upper bounded by a positive vector $\bar{\mathbf{g}} \in \Re^{|\mathcal{C}|}$, that is, $\mathbf{g}(\boldsymbol{\pi}) < \bar{\mathbf{g}}$ holds for every $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$. Then, it is possible to construct a *fixed* demand model which is equivalent to the elastic demand Wardrop conditions (2). The model is constructed as follows. For each OD pair $(p, q) \in \mathcal{C}$, we introduce an additional link from node p to node q which carries a nonnegative flow e_{pq}, and which has as its link cost $\xi_{pq}(\bar{\mathbf{g}} - \mathbf{e})$, $\mathbf{e} \in \Re^{|\mathcal{C}|}$. This link carries a flow which is the difference between the upper bound $\bar{g}_{pq}$ and the flow $\sum_{r \in \mathcal{R}_{pq}} h_r$ on that OD pair in the *original* network; hence, $\boldsymbol{\Gamma}^\top \mathbf{h} + \mathbf{e} = \bar{\mathbf{g}}$ holds. We call this new link the *excess demand* link for OD pair (p, q). See Figure 2 for an illustration.

The form of this fixed demand model is such that, at equilibrium, the excess demand link carries some flow. Therefore, the costs on all used routes on that OD pair must match the value of the inverse demand function for that OD pair, and the flow in the original network equals that specified by the demand function on a least-cost route. We conclude that the fixed demand equilibrium in the modified network is equivalent to an elastic demand in the original network. The complete fixed demand model consists in finding flows $(\mathbf{h}, \mathbf{e})$ such that for some vector $\boldsymbol{\pi} \in \Re^{|\mathcal{C}|}$,

$$\mathbf{0}^{|\mathcal{R}|} \leqslant \mathbf{h} \perp \left(\mathbf{c}(\mathbf{h}) - \boldsymbol{\Gamma}\boldsymbol{\pi}\right) \geqslant \mathbf{0}^{|\mathcal{R}|}, \tag{10a}$$

$$\mathbf{0}^{|\mathcal{C}|} \leqslant \mathbf{e} \perp \left(\boldsymbol{\xi}(\bar{\mathbf{g}} - \mathbf{e}) - \boldsymbol{\pi}\right) \geqslant \mathbf{0}^{|\mathcal{C}|}, \tag{10b}$$

$$\boldsymbol{\Gamma}^\top \mathbf{h} + \mathbf{e} = \bar{\mathbf{g}}. \tag{10c}$$

We remark that this model is equivalent to a fixed demand VIP of the form

$$-\begin{pmatrix} \mathbf{c}(\mathbf{h}) \\ \boldsymbol{\xi}(\bar{\mathbf{g}} - \mathbf{e}) \end{pmatrix} \in N_{H_{\bar{g}}}(\mathbf{h}, \mathbf{e}), \tag{11a}$$

where

$$H_{\bar{g}} := \left\{(\mathbf{h}, \mathbf{e}) \in \Re_+^{|\mathcal{R}|} \times \Re_+^{|\mathcal{C}|} \middle| \boldsymbol{\Gamma}^\top \mathbf{h} + \mathbf{e} = \bar{\mathbf{g}}\right\}. \tag{11b}$$

We further note that the excess demand cost

$$\boldsymbol{\kappa}(\mathbf{e}) := \boldsymbol{\xi}(\bar{\mathbf{g}} - \mathbf{e}), \tag{12}$$

whose domain is the open box $\{\mathbf{e} \in \Re^{|\mathcal{C}|} \mid \mathbf{0}^{|\mathcal{C}|} < \mathbf{e} < \bar{\mathbf{g}}\}$, is strictly monotone if and only if $-\mathbf{g}$ is. For this model, we have the following result.

Proposition 1 (Equivalent fixed demand model). *Suppose that the function* $\mathbf{g} : \Re^{|\mathcal{C}|} \to \Re^{|\mathcal{C}|}$ *is invertible, with inverse* $\boldsymbol{\xi} : \Re^{|\mathcal{C}|} \to \Re^{|\mathcal{C}|}$. *Suppose further that* $\mathbf{g}$ *is upper bounded by a positive demand vector* $\bar{\mathbf{g}} \in \Re^{|\mathcal{C}|}$. *Then, the fixed demand model* (10) *is equivalent to the elastic demand model* (2).

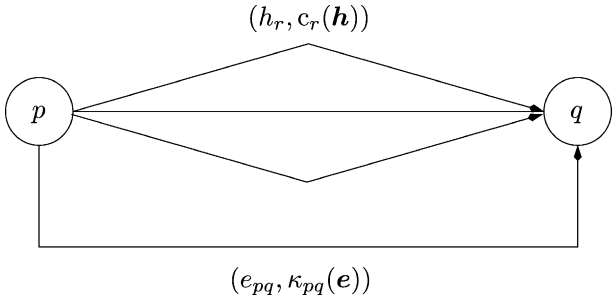


Fig. 2. The excess demand link for OD pair (p, q).

2.5 *The equivalence between equilibria in different representations*

The models discussed in Section 2.3 are distinct. The difference between the models VIP($\mathbf{t}$, $\widehat{F}$) (respectively, VIP($[\mathbf{t}, -\boldsymbol{\xi}]$, $\widehat{F}_d$)) and VIP($\mathbf{t}$, F) (respectively, VIP($[\mathbf{t}, -\boldsymbol{\xi}]$, F_d)) lies in possible occurrence of *cycles* in the latter's admissible sets. (Note that the definition of the sets $\mathcal{R}_{pq}$ specifies that routes are cycle-free.) We therefore have that $\widehat{F} \subseteq F$.[3] The equivalence between the link–route and link–node representations in terms of equilibria thus hinges on whether an equilibrium flow in the latter can contain a cycle; we provide such a result below. The interest in this topic is not only theoretical, as a poor modeling clearly can lead to spurious results; normally, in traffic networks, no one would travel in a cycle, but the improper modeling of a traffic equilibrium problem may lead to such a result in theory, when using the F (or F_d) representation of flows. We note that no cycles are present in the numerical example.

Since the demand part of the model plays no role in the analysis, we can focus on the fixed demand case. The result below carries over trivially to elastic demand models involving bounded demand functions. The proof rests on the following flow decomposition result.

Theorem 2 (Flow Decomposition Theorem). *Every route and cycle flow has a unique representation as nonnegative link flows. Conversely, every nonnegative link flow may be represented as a route and cycle flow (though not necessarily uniquely), which utilizes at most $|\mathcal{L}| + |\mathcal{N}|$ routes and cycles, and of which at most $|\mathcal{N}|$ are cycles.*

We say that a link flow $\mathbf{v} \in F$ is *weakly acyclic* if there exists a decomposition into commodity link flows $\mathbf{w}_k$ that does not use all links of a directed cycle. The flow is said to be *strongly acyclic* if there exists no decomposition into commodity link flows $\mathbf{w}_k$ that use all the links of a directed cycle. The following

[3] The "$\widehat{\cdot}$" notation reflects that the set $\widehat{F}$ is bounded, while F may be unbounded.

relationships are immediate from the Flow Decomposition Theorem:

$$\{\mathbf{v} \in F \mid \mathbf{v} \text{ is strongly acyclic}\} \subseteq \{\mathbf{v} \in F \mid \mathbf{v} \text{ is weakly acyclic}\} \subseteq \widehat{F} \subseteq F.$$

The decomposition of a link flow solution can of course be performed with respect to other entities than routes and cycles, as was done in the Flow Decomposition Theorem 2. For example, for a fixed link flow vector $\mathbf{v} \in F$, consider the linear program to

$$\underset{(\hat{\mathbf{v}},\boldsymbol{\delta})}{\text{minimize}}\ \mathbf{1}^\top \hat{\mathbf{v}}, \tag{13a}$$

$$\text{subject to} \quad \hat{\mathbf{v}} \in F, \tag{13b}$$

$$\hat{\mathbf{v}} + \boldsymbol{\delta} = \mathbf{v}, \tag{13c}$$

$$\boldsymbol{\delta} \geqslant \mathbf{0}^{|\mathcal{L}|}. \tag{13d}$$

The result of this linear program is a decomposition of the link flow $\mathbf{v}$ into a sum $\hat{\mathbf{v}} + \boldsymbol{\delta}$ of a *strongly acyclic* link flow $\hat{\mathbf{v}}$, and a *circulation* link flow $\boldsymbol{\delta}$, that is, a flow which satisfies $\mathbf{E}\boldsymbol{\delta} = \mathbf{0}^{|\mathcal{N}|}$. Moreover, $\mathbf{v}$ is strongly acyclic if and only if $\mathbf{v} = \hat{\mathbf{v}}$. This linear program, thus, also provides a decomposition of a link flow in terms of extreme points and directions of the individual flow sets F_{pq}, although not disaggregated into routes and cycles in the individual commodity spaces. It is therefore different from the representation described in the Flow Decomposition Theorem, as there may be a circulation in $\hat{\mathbf{v}}$, although such a flow must necessarily include more than one OD pair.

It follows from the above analysis that one has to be careful when implementing algorithms based on link (vs. route) flows. Indeed, convex combinations of cycle-free flows may fail to be cycle-free, unless restrictive assumptions hold.

Finally, we say that the link cost vector $\mathbf{t}$ is *cycle-wise nonnegative* (respectively, *cycle-wise positive*) on F if the sum of the link costs of every cycle is nonnegative (respectively, positive) on F.

Proposition 3 (Equilibria for two representations).

(a) *If* $\mathbf{t}$ *is cycle-wise nonnegative on* F*, then every solution to* VIP($\mathbf{t}, \widehat{F}$) *is also a solution to* VIP($\mathbf{t}, F$).

(b) *Every weakly acyclic solution to* VIP($\mathbf{t}, F$) *solves* VIP($\mathbf{t}, \widehat{F}$).

(c) *If* $\mathbf{t}$ *is cycle-wise positive on* F*, then every solution to* VIP($\mathbf{t}, F$) *is strongly acyclic.*

Clearly, then, if link costs are everywhere positive the issue of possible cyclic equilibrium flows is avoided. One must however note that when considering general link toll models, zero and even negative link costs may occur, as they may appeal to certain travelers.

2.6 Reduction to an optimization problem

In the case where $\mathbf{t}$ is a gradient mapping,[4] the model VIP($\mathbf{t}$, $\widehat{F}$) defines the first-order optimality conditions for an optimization problem of the form

$$\underset{\mathbf{v}\in\widehat{F}}{\text{minimize}}\ \phi(\mathbf{v}) := \oint_{\mathbf{0}^{|\mathcal{L}|}}^{\mathbf{v}} \mathbf{t}(s)\,\mathrm{d}s, \tag{14}$$

where $\oint$ denotes a line integral. The equivalence between equilibria, that is, solutions to the model VIP($\mathbf{t}$, $\widehat{F}$), and stationary points to the problem (14) follows immediately from the fact that $\nabla\phi \equiv \mathbf{t}$. In particular, every stationary point of problem (14) is an equilibrium link flow.

Further, if $\mathbf{t}$ is separable, so that t_l is a function only of v_l, $l \in \mathcal{L}$, the optimization problem assumes the more familiar form

$$\underset{\mathbf{v}\in\widehat{F}}{\text{minimize}}\ \phi(\mathbf{v}) := \sum_{l\in\mathcal{L}} \int_0^{v_l} t_l(s)\,\mathrm{d}s. \tag{15}$$

When the demand function is separable and invertible, that is, $g_{pq}^{-1} = \xi_{pq}$, we similarly have the equivalent optimization problem

$$\underset{(\mathbf{v},\mathbf{d})\in\widehat{F}_d}{\text{minimize}}\ \phi(\mathbf{v},\mathbf{d}) := \sum_{l\in\mathcal{L}} \int_0^{v_l} t_l(s)\,\mathrm{d}s - \sum_{(p,q)\in\mathcal{C}} \int_0^{d_{pq}} \xi_{pq}(s)\,\mathrm{d}s. \tag{16}$$

It follows as above from the identity $\nabla\phi \equiv (\mathbf{t}, -\boldsymbol{\xi})$ that elastic demand equilibria are stationary points for (16).

Since the cost and demand mappings are identical irrespective of the flow representation, the traffic equilibrium models for the link–node flow representation correspond to optimization models that have the same form as above, except that $\widehat{F}$ and $\widehat{F}_d$ are replaced, respectively, by F and F_d. The reader may provide the corresponding optimization formulation to the above numerical example, and show that the equilibrium solution satisfies the first-order optimality conditions associated with the mathematical program (16).

2.7 Properties of equilibria

In this section we derive the basic properties of traffic equilibria, i.e., the existence and uniqueness of the following entities: flows, demands, and equilibrium travel costs.[5]

[4] If $\mathbf{t}$ is continuously differentiable relative to an open convex set containing S, then $\mathbf{t}$ is a gradient mapping over S if and only if its Jacobian matrix $\nabla\mathbf{t}(\mathbf{v})$ is symmetric over S. The gradient property is a more general property than this symmetry property, since $\mathbf{t}$ need not always be differentiable; nevertheless, the term 'symmetry' is frequently used in traffic planning when it should really refer to the gradient property, also called the 'integrability' property.

[5] For ease of presentation, we work with additive link costs throughout although, with regards to the existence result in Theorem 4(a), this condition is not necessary.

We state the properties of the models that hold throughout this section:

(i) The network $\mathcal{G}$ is strongly connected with respect to the OD pairs[6];

(ii) the function $\mathbf{t}: \Re_+^{|\mathcal{L}|} \to \Re^{|\mathcal{L}|}$ is single-valued and continuous on $\Re_+^{|\mathcal{L}|}$;

(iii) the function $\mathbf{g}: \Re^{|\mathcal{C}|} \to \Re^{|\mathcal{C}|}$ is single-valued, continuous, nonnegative and upper bounded on $\Re^{|\mathcal{C}|}$.

2.7.1 Existence of equilibria

Theorem 4 (Existence of an equilibrium).

(a) Link–route representation. *There exists a compact set of vectors* $(\mathbf{h}, \mathbf{d}, \boldsymbol{\pi})$ *of flows, demands, and least travel costs that solve the traffic equilibrium model* (2).

(b) Link–node representation. *Suppose further that the function* $\mathbf{c}$ *is additive, and that the function* $\mathbf{t}$ *is cycle-wise nonnegative. Then, there exists a compact set of vectors* $((\mathbf{w}_k)_{k\in\mathcal{C}}, \mathbf{v}, \mathbf{d}, (\boldsymbol{\pi}_k)_{k\in\mathcal{C}})$ *of flows, demands, and node prices, solving the traffic equilibrium model* (6). *Moreover, the link flows* $((\mathbf{w}_k)_{k\in\mathcal{C}}, \mathbf{v})$ *can be taken to be strongly acyclic.*

Proof.

(a) The proof hinges on a general existence result for variational inequalities, which we will utilize by converting the traffic equilibrium model into an equivalent VIP involving a continuous operator defined over a nonempty, convex and *compact* set. Since $\mathbf{g}$ is nonnegative and upper bounded, the variable $\mathbf{h}$ resides in a bounded subset of the nonnegative orthant; since $\mathbf{c}$ is continuous, the variable $\boldsymbol{\pi}$ resides in a compact subset of $\Re^{|\mathcal{C}|}$, although not necessarily in the nonnegative orthant. Hence, the variables of the problem, $(\boldsymbol{\pi}, \mathbf{h})$, lie in a compact set. We can therefore, with no loss of generality, introduce additional, redundant, bounds for each variable vector, such that

$$\underline{\boldsymbol{\pi}} \leqslant \boldsymbol{\pi} \leqslant \bar{\boldsymbol{\pi}},$$
$$\mathbf{h} \leqslant \bar{\mathbf{h}},$$

where $\underline{\boldsymbol{\pi}} \leqslant \bar{\boldsymbol{\pi}} \in \Re^{|\mathcal{C}|}$ are small and large enough, respectively, and $\bar{\mathbf{h}} \in \Re^{|\mathcal{R}|}$ is large enough. Incorporating these constraints yields a VIP model for (2) that involves a continuous operator defined over a nonempty, convex and compact set. Hence, its solution set is nonempty and compact.[7] Furthermore, the bounds can clearly be selected such that they all are fulfilled strictly at any equilibrium solution.

[6] This condition means that there exists at least one route between each OD pair.

[7] The idea is to construct, by means of the projection operator, a fixed point problem equivalent to the original VIP, and then to apply Brouwer's fixed point theorem.

(b) Suppose that the vectors $(\mathbf{h}, \boldsymbol{\pi})$ and $\mathbf{d} := \mathbf{g}(\boldsymbol{\pi})$ solve (2), and set $\mathbf{v} := \boldsymbol{\Lambda}\mathbf{h}$; the existence of a solution follows from the result in (a) above. Then, the vector $\mathbf{v}$ solves the fixed demand traffic equilibrium model VIP($\mathbf{t}, \widehat{F}$). We now make use of Proposition 3(a), to conclude that $\mathbf{v}$ also solves VIP($\mathbf{t}, F$). We next need to establish that this flow is consistent with the elastic demand model (6). To this end, we define the $|\mathcal{N}|$-vectors $\mathbf{d}_k$ and $\boldsymbol{\pi}_k$, $k \in \mathcal{C}$, by the use of the incidence vectors $\mathbf{i}_k$, $k \in \mathcal{C}$, discussed in Section 2.2, and note that the vectors $\boldsymbol{\pi}_k$, $k \in \mathcal{C}$ are consistent with the vector $\boldsymbol{\pi}$. The first result then follows.

Last, we establish that the link flow $\mathbf{v}$ can be taken to be strongly acyclic. Consider the perturbed link cost operator $\mathbf{t} + \varepsilon_\tau \cdot \mathbf{1}^{|\mathcal{L}|}$, $\varepsilon_\tau > 0$. This mapping is cycle-wise positive, so Proposition 3(c) states that every equilibrium link flow is strongly acyclic. Reasoning as above, and letting $\{\varepsilon_\tau\}$ tend to zero as τ tends to infinity, we obtain the desired results, since the set of strongly acyclic flows is compact. □

We note that since a fixed demand vector is a special case of the function $\mathbf{g}$ stated previously, the above result provides an existence result in both the elastic and fixed demand cases.

2.7.2 *Uniqueness of equilibria*

To establish the uniqueness of the equilibrium solution is advantageous from several perspectives, but most of all from that of practice: using an equilibrium model to make predictions about entities that are not unique at equilibrium is a hazardous at best. Secondly, uniqueness is required for the convergence of some algorithms that seek an equilibrium. Thirdly, when traffic equilibrium is but a submodel in a more general model, uniqueness is crucial in order for the overall model to be solvable, or even consistent. For example, a sensitivity analysis is impossible to perform in entities that are not properly defined.

While the existence of an equilibrium was ensured under conditions that are almost identical for all of the models that we have formulated, this is not the case with the uniqueness issue. For example, there is a clear distinction between uniqueness results for variational inequality models and optimization models, when the cost mapping is not integrable. Further, uniqueness is a property that depends on the level of aggregation. For example, route flows (or, commodity link flows) are almost never unique at equilibrium, at least when travel costs are additive. The reason is of course that a link flow $\mathbf{v}$ almost always can be distributed across the routes (or, disaggregated over the commodities on the links) in infinitely many ways into a route flow $\mathbf{h}$ while the relation $\mathbf{v} = \boldsymbol{\Lambda}\mathbf{h}$ is satisfied (or, similarly, into an infinite variety of OD link flows $\mathbf{w}_k$ satisfying $\sum_{k \in \mathcal{C}} \mathbf{w}_k = \mathbf{v}$). In the results to follow, we will see that the uniqueness of the entities $\mathbf{c}$ (and $\mathbf{t}$ if present), $\boldsymbol{\pi}$, and $\mathbf{d}$ are ensured under the mildest conditions. In contrast, stronger conditions are required to ensure the uniqueness of *even* the link flow vector $\mathbf{v}$.

The results below are provided for the link–route representation only. Indeed, provided that the conditions of Theorem 4(b) hold, the assumptions that imply uniqueness are the same for both representations. The theorem is stated under fairly weak conditions that involve technical concepts that are detailed in Appendix A.

Theorem 5 (Uniqueness of the equilibrium: The general case).

(a) Convexity. *Suppose that* $(\mathbf{c}, -\mathbf{g})$ *is monotone. Then, the sets of equilibrium route and link flows* $(\mathbf{h}, \mathbf{v})$*, least travel costs* $\boldsymbol{\pi}$*, and travel demands* $\mathbf{d}$*, are convex. Suppose further that the travel demand is fixed. Then, the same conclusion holds if the monotonicity requirement on* $\mathbf{c}$ *is replaced by the weaker pseudo-monotonicity condition.*

(b) Travel costs and demands. *Suppose that* $(\mathbf{c}, -\mathbf{g})$ *is monotone*$^+$*. Then both the functions* $\mathbf{c}$ *and* $\mathbf{g}$ *are constant over the set of equilibria. This implies that both the route costs* $\mathbf{c}(\mathbf{h})$*, least travel costs* $\boldsymbol{\pi}$*, and travel demands* $\mathbf{d}$*, are unique. If* $\mathbf{c}$ *is additive, and* $\mathbf{t}$ *is monotone*$^+$*, then* $\mathbf{t}(\mathbf{v})$ *is unique on the set of equilibria as well.*

If further, the travel demand function is constant, the same conclusion holds for the entities $\mathbf{c}(\mathbf{h})$ *and* $\boldsymbol{\pi}$ *under the weaker requirement that* $\mathbf{c}$ *is pseudomonotone*$^+_*$*. If* $\mathbf{c}$ *is additive and* $\mathbf{t}$ *is pseudomonotone*$^+_*$*, then* $\mathbf{t}(\mathbf{v})$ *is unique on the set of equilibria as well.*

(c) Link flows. *Suppose that* $(\mathbf{c}, -\mathbf{g})$ *is monotone, that* $\mathbf{c}$ *is additive, and that* $\mathbf{t}$ *is strictly monotone. Then, the equilibrium link flow* $\mathbf{v}$ *is unique, as well as* $\mathbf{c}$*,* $\boldsymbol{\pi}$*, and* $\mathbf{d}$*, and the set of equilibrium route flows is a polytope.*

Proof (Sketch). (a) That the solution set is nonempty and compact follows from the existence Theorem 4. The convexity of the solution set and its entities is a consequence of pseudomonotonicity and the properties of projections onto convex sets. The reason why pseudomonotonicity is not enough for convexity to hold in the elastic demand case is that pseudomonotonicity is not preserved under addition.

(b) This result follows from a general one for the variational inequality problem VIP$(\mathbf{f}, X)$: if $\mathbf{f}$ is pseudomonotone$^+_*$ on X, then the value $\mathbf{f}(\mathbf{x})$ is constant on SOL$(\mathbf{f}, X)$. Again, the property pseudomonotone$^+_*$ must be replaced by monotone$^+$ for the elastic demand case, because the former property is not preserved under addition.

(c) Follows easily by arguing through contradiction. □

The result for the separable case follows as a corollary, because the property of monotonicity$_+$ of a mapping is identical to that of monotonicity when it is integrable, and pseudomonotonicity$^+_*$ reduces to pseudomonotonicity for functions defined over $\Re$.

Corollary 6 (Uniqueness of the equilibrium: The separable case).

(a) Travel costs and demands. *Suppose that each function* g_{pq}, $(p, q) \in \mathcal{C}$, *is decreasing, and that each function* t_l, $l \in \mathcal{L}$, *is increasing. Then, the route and link costs* $\mathbf{c}$ *and* $\mathbf{t}$, *least travel costs* $\boldsymbol{\pi}$, *and travel demands* $\mathbf{d}$, *are unique. In particular, if each function* g_{pq}, $(p, q) \in \mathcal{C}$, *is invertible, then it is strictly decreasing, and on the set of optimal solutions for the convex optimization formulation* (16), *the equilibrium link travel costs* $\mathbf{t}(\mathbf{v})$, *least costs* $\boldsymbol{\pi}$, *and demands* $\mathbf{d}$ *are unique.*

Suppose further that the travel demand is fixed. Then, the same conclusion holds for the entities $\mathbf{c}$, $\mathbf{t}$, *and* $\boldsymbol{\pi}$ *on the set of optimal solutions to the convex optimization formulation* (15) *if the monotonicity requirement on* t_l *is replaced by pseudomonotonicity.*

(b) Link flows. *Suppose that each function* t_l, $l \in \mathcal{L}$, *is strictly increasing. Then, in the solutions to the optimization formulations* (15) *and* (16), *the equilibrium link flow* $\mathbf{v}$ *is unique.*

To summarize in a more convenient and compact form, we have for the elastic demand case:

$$\begin{aligned}
&\mathbf{t} \text{ monotone}, -\mathbf{g} \text{ strictly monotone} \quad \Longrightarrow \quad \boldsymbol{\pi}, \mathbf{d} \text{ unique},\\
&(\mathbf{t}, -\mathbf{g}) \text{ monotone}^{+} \quad \Longrightarrow \quad \mathbf{c}(\mathbf{h}), \mathbf{t}(\mathbf{v}), \boldsymbol{\pi}, \mathbf{d} \text{ unique},\\
&\mathbf{t} \text{ strictly monotone}, -\mathbf{g} \text{ monotone} \quad \Longrightarrow \quad \mathbf{v}, \mathbf{c}(\mathbf{h}), \mathbf{t}(\mathbf{v}), \boldsymbol{\pi}, \mathbf{d} \text{ unique}.
\end{aligned}$$

The first, and slightly weaker, of these results is not established above, but follows easily by arguing through contradiction. For fixed demands, the corresponding results are:

$$\begin{aligned}
&\mathbf{t} \text{ pseudomonotone}^{+}_{*} \quad \Longrightarrow \quad \mathbf{c}(\mathbf{h}), \mathbf{t}(\mathbf{v}), \boldsymbol{\pi} \text{ unique},\\
&\mathbf{t} \text{ strictly monotone} \quad \Longrightarrow \quad \mathbf{v}, \mathbf{c}(\mathbf{h}), \mathbf{t}(\mathbf{v}), \boldsymbol{\pi} \text{ unique}.
\end{aligned}$$

2.8 Alternative notions of traffic equilibria and their relations

In this section, we present a few equilibrium conditions that have emerged since the original one defined by Wardrop, and relate them to his. Some of the motivation for this development is that when modeling the effect of congestion pricing instruments, the cost term associated with the pricing scheme might take the form of a (discontinuous) step function; in order to be able to study the existence and computations of equilibria for such discontinuous travel costs, we require that the traditional equilibrium conditions be generalized to allow for certain types of discontinuities.

Alternative definitions

We consider (without any loss of generality) a fixed demand traffic network $\mathcal{G} = (\mathcal{N}, \mathcal{L})$. In this network, we assume that the traffic volume $\mathbf{h} \in H$ can be observed. We further focus on an arbitrary OD pair $(p, q) \in \mathcal{C}$. More precisely,

we consider a route $r \in \mathcal{R}_{pq}$ for which $h_r > 0$. We also consider a cost function $\mathbf{c}: \Re_+^{|\mathcal{R}|} \to \Re^{|\mathcal{R}|}$, each component of which $c_r: \Re_+^{|\mathcal{R}|} \to \Re$ is assumed to be single-valued. Last, we consider a vector $\mathbf{p}$ in $\Re^{|\mathcal{R}|}$ describing the move of one unit of traffic volume from route $r \in \mathcal{R}_{pq}$ to route s across the same OD pair, that is,

$$p_i = \begin{cases} -1, & \text{if } i = r, \\ 1, & \text{if } i = s, \\ 0, & \text{otherwise.} \end{cases}$$

The following definitions summarize the equilibrium conditions considered in this section. Notice that we here distinguish the terms 'Wardrop equilibrium' and 'user equilibrium', although they are frequently used interchangeably in this text. These notions coincide whenever the travel cost is continuous, as established below, whence there is no confusion in most situations.

Definition 7 (Equilibrium conditions).

(a) Wardrop equilibrium. The traffic volume $\mathbf{h}$ is a Wardrop equilibrium if

$$c_r(\mathbf{h}) \leqslant c_s(\mathbf{h}). \tag{17}$$

The traffic volume is a Wardrop equilibrium if for each driver, the present travel cost on any alternative route is at least as high as the present one on its present route.

(b) User optimized. The traffic volume $\mathbf{h}$ is user optimized if for some $\alpha > 0$,

$$c_r(\mathbf{h}) \leqslant c_s(\mathbf{h} + \varepsilon\mathbf{p}), \quad \varepsilon \in \big[0, \text{minimum}\{\alpha, h_r\}\big]. \tag{18}$$

The traffic volume is user optimized if any driver who switches routes experiences a travel cost that is at least as high as the present one on its present route.

(c) Equilibrated. The traffic volume $\mathbf{h}$ is equilibrated if

$$c_r(\mathbf{h} + \varepsilon\mathbf{p}) \leqslant c_s(\mathbf{h} + \varepsilon\mathbf{p}), \quad \varepsilon \in [0, h_r]. \tag{19}$$

The traffic volume is equilibrated if any driver who switches routes experiences a travel cost that is at least as high as the new one on its present route.

(d) User equilibrium. The traffic volume $\mathbf{h}$ is a user equilibrium if

$$c_r(\mathbf{h}) \leqslant \liminf_{\varepsilon \downarrow 0} c_s(\mathbf{h} + \varepsilon\mathbf{p}). \tag{20}$$

The traffic volume is a user equilibrium if no arbitrarily small packet of flow can reroute itself and lower its travel cost.

(e) Normal equilibrium. The traffic volume $\mathbf{h}$ is a normal equilibrium if it satisfies

$$\mathbf{c}(\mathbf{y})^\top(\mathbf{y} - \mathbf{h}) \geqslant 0, \quad \mathbf{y} \in H. \tag{21}$$

The definition states that $\mathbf{h} \in H$ solves the Minty variational inequality associated with the traffic system.

Relationships

In the following, we show how the above definitions are related to each other. We assume here that the travel costs are additive, that is, $c_r(\mathbf{h}) = \sum_{l \in \mathcal{L}} \lambda_{lr} t_l(\mathbf{v})$ for every consistent pair $(\mathbf{h}, \mathbf{v})$ of route and link volumes. Note that the l.s.c. and u.s.c. properties[8] are additive, so if link travel costs are l.s.c. (respectively, u.s.c.) then the route costs are l.s.c. (respectively, u.s.c.) as well.

Proposition 8 (Relationships between equilibrium definitions). *Let the route travel cost* $\mathbf{c}$ *be additive.*

(a) Wardrop equilibria vs. user optimized flows. *Suppose that the route cost function* $\mathbf{c}$ *is continuous. Then, user optimized flows are Wardrop equilibria. If the route cost function* $\mathbf{c}$ *is in* C^1 *on* H *and, for every* $\mathbf{h} \in H$ *and routes* $r, s \in \mathcal{R}_{pq}$, $(p, q) \in \mathcal{C}$, *the inequality*

$$\frac{\partial c_s(\mathbf{h})}{\partial h_s} \geqslant \frac{\partial c_s(\mathbf{h})}{\partial h_r} \tag{22}$$

holds, Wardrop equilibria are user optimized flows. In particular, if the link cost function $\mathbf{t}$ *is separable and increasing, Wardrop equilibria are user optimized.*

(b) Wardrop equilibria vs. user equilibrated flows. *Suppose that the route cost function* $\mathbf{c}$ *is continuous. Then, equilibrated flows are Wardrop equilibria. If the route cost function* $\mathbf{c}$ *is pseudomonotone on* H, *Wardrop equilibria are equilibrated flows.*

(c) Wardrop equilibria vs. user equilibria. *Suppose that for each link* $l \in \mathcal{L}$, t_l *is l.s.c. Then, Wardrop equilibria are user equilibria. If, for each link* $l \in \mathcal{L}$, t_l *is u.s.c., then user equilibria are Wardrop equilibria. Whenever the link cost is continuous, the concepts of Wardrop and user equilibria coincide.*

(d) Wardrop equilibria vs. normal equilibria. *Suppose that the link cost is l.s.c. Then, normal equilibria are Wardrop equilibria. If the link cost operator* $\mathbf{t}$ *is pseudomonotone, Wardrop equilibria are normal equilibria.*

(e) User equilibria vs. normal equilibria. *Suppose that the link cost is l.s.c. Then, normal equilibria are user equilibria.*

Proof. In the case of (b), that Wardrop equilibria are equilibrated flows, suppose that $\mathbf{c}$ is pseudomonotone on H, and let $\mathbf{h} \in H$ be a Wardrop equilibrium.

[8] A function $\phi : \Re^n \to \Re \cup \{-\infty, +\infty\}$ is *lower semicontinuous* (*l.s.c.*) at $\mathbf{x} \in \Re^n$ if $\phi(\mathbf{x}) = \liminf_{\mathbf{y} \to \mathbf{x}} \phi(\mathbf{y})$, $\mathbf{x} \in \Re^n$, and *upper semicontinuous* (*u.s.c.*) at $\mathbf{x}$ if $\phi(\mathbf{x}) = \limsup_{\mathbf{y} \to \mathbf{x}} \phi(\mathbf{y})$, $\mathbf{x} \in \Re^n$. A function ϕ is l.s.c. (respectively, u.s.c.) on the set $S \subseteq \Re^n$ if it is l.s.c. (respectively, u.s.c.) at every point $\mathbf{x} \in S$. The function ϕ is *continuous* at $\mathbf{x}$ if it is both l.s.c. and u.s.c. at $\mathbf{x}$.

Then, by definition and for all $\varepsilon > 0$ with $\mathbf{h} + \varepsilon\mathbf{p} \in H$, there holds

$$0 \leqslant \mathbf{c}(\mathbf{h} + \varepsilon\mathbf{p})^\top \big([\mathbf{h} + \varepsilon\mathbf{p}] - \mathbf{h}\big) = \varepsilon\mathbf{c}(\mathbf{h} + \varepsilon\mathbf{p})^\top\mathbf{p}$$
$$= \varepsilon\big[c_s(\mathbf{h} + \varepsilon\mathbf{p}) - c_r(\mathbf{h} + \varepsilon\mathbf{p})\big],$$

where the inequality follows since $\mathbf{h} \in H$ is a Wardrop equilibrium and $\mathbf{c}$ is pseudomonotone on H. As $\varepsilon > 0$, we are done.

In the case of (d), the result is established by referring to the equivalence (under the assumption of pseudomonotonicity of $\mathbf{t}$) between the variational inequality problem VIP($\mathbf{c}, H$) and the Minty formulation (21); see also (104). In the proof of the first part (that solutions to the Minty variational inequality are solutions to the standard variational inequality problem), continuity can be replaced by lower semicontinuity. The reverse implication utilizes the notion of pseudomonotonicity. For the other results, we refer to the Notes section. □

Remark 9 (On the relationships between the equilibrium conditions). The matrix condition (22) appearing in (a) describes a restriction on the dependence between costs on distinct routes. In broad terms, if routes that are alternatives to each other do not interact too strongly, then Wardrop equilibria are user optimized. Differentiable, separable and increasing link travel cost functions satisfy the matrix condition (22).

Under weak conditions, viz. continuity and monotonicity, four of the concepts collapse to Wardrop's definition of user equilibrium. The outlier is the user optimized concept, akin to a nonatomic Nash equilibrium solution.[9] The following counter-example shows that a Wardrop equilibrium may not be user optimized when the travel cost is asymmetric.

Consider a two-node network with parallel routes (links) and with respective costs

$$c_1(h_1, h_2) = 2h_1 + 3h_2,$$
$$c_2(h_1, h_2) = 2h_2 + 25.$$

Note that this mapping is monotone. If demand is equal to 15, the flow vector $\mathbf{h} = (10, 5)$ is an equilibrium, with common path cost 34. Now, let us shift one flow unit from the second to the first path ($\delta = 1$). The cost vector of the new flow pattern $\mathbf{h}' = (11, 4)$ is $c(h'_1, h'_2) = (34, 33)$. Since the cost to users that have switched path has decreased,[10] the initial flow $\mathbf{h}$ is not a Nash equilibrium. Actually, there exist neither user optimized (nor Nash) equilibria for this example.

The relationships between the various concepts are illustrated in Figure 3.

[9] A Nash equilibrium between finitely many players is achieved when no player, acting on its own, can improve its payoff. A nonatomic game involves a continuum of players.

[10] In fact, all user costs have decreased.

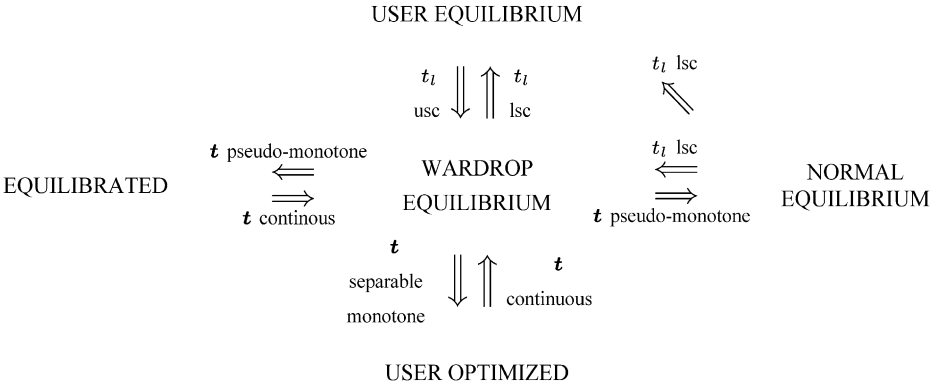


Fig. 3. Relations among definitions of equilibria.

2.9 *User versus system optimality*

In his seminal paper, Wardrop introduced the concepts of user and system equilibrium, the second depicting flows that minimize the total system cost $S(\mathbf{v}) := \mathbf{t}(\mathbf{v})^\top \mathbf{v}$.

A question that naturally arises is: how different are equilibrium and system-optimal flows? How much would be gained, in terms of efficiency, from a centralized control of all traffic flow? Looking at the question from a worst-case perspective, we want to determine

$$r(\Delta) = \underset{\Delta}{\text{maximum}} \frac{\mathbf{t}(\mathbf{v}_{\text{UO}})^\top \mathbf{v}_{\text{UO}}}{\mathbf{t}(\mathbf{v}_{\text{SO}})^\top \mathbf{v}_{\text{SO}}}, \tag{23}$$

where $\mathbf{v}_{\text{SO}}$ (respectively $\mathbf{v}_{\text{UO}}$) denotes a system-optimal (respectively user-optimal) flow pattern, and Δ regroups the data of the equilibrium problem, that is, $\Delta = \mathcal{G} \cup \mathbf{g} \cup \mathcal{C} \cup \mathbf{t}$ in the fixed demand case. This ratio $r(\Delta)$ is the *price of anarchy* and, in order that it be well defined, we need that the system cost associated with user equilibria be unique. This condition will be satisfied under the monotonicity requirements outlined in Theorem 5.

In certain circumstances, for instance if the cost function is link-separable and assumes the monomial form

$$t_l(v_l) = \kappa_l v_l^\mu, \quad \kappa_l, \mu \geqslant 0,\ l \in \mathcal{L},$$

user and system-optimal flows coincide. This is not the case in general, as can be observed in the example of Figure 4, where $\mathbf{v}_{\text{UO}}^\top = (v_{AB}, v_{AC}, v_{BC}, v_{BD}, v_{CD})^\top = (4, 2, 2, 2, 4)$[11] and $\mathbf{v}_{\text{SO}}^\top = (3, 3, 0, 3, 3)$. In this situation, every user is worse off at (user) equilibrium, and the price of anarchy is equal to 92/83,

[11] This corresponds to assigning flow in equal proportions on each of the three routes of the network.

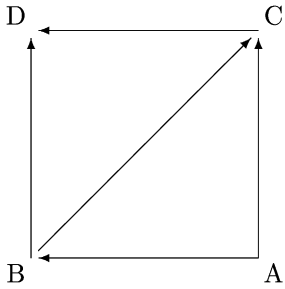


Fig. 4. The Braess paradox network.

as can readily be computed. The following theorem provides an exact value of $r(\Delta)$ for the important class of polynomial cost functions.

Theorem 10 (The price of anarchy). *If the mapping* $\mathbf{t}$ *is link-separable, increasing and polynomial of degree* p*, the price of anarchy is equal to*

$$\frac{(p+1)^{1+1/p}}{(p+1)^{1+1/p} - p} \in \Theta\left(\frac{p}{\ln p}\right).$$

If, further, $\mathbf{t}$ *is affine and monotone* ($p = 1$)*, then the price of anarchy is equal to* 4/3.

Let us now define the *steepness* of a cost mapping $\mathbf{t}$ as

$$\sigma(\mathbf{t}) = \sup_{l \in \mathcal{L}, \mathbf{v} \in F} \frac{\mathbf{t}(\mathbf{v})^\top v}{\mathbf{t}(\mathbf{v})^\top \mathbf{v} + \gamma_l(\mathbf{v}) t_l(\gamma_l(\mathbf{v})) - \gamma_l(\mathbf{v}) t_l(\mathbf{v})}, \tag{24}$$

where β is a function such that $\nabla S(\beta(\mathbf{v})) = \mathbf{t}(\mathbf{v})$.

Theorem 11 (The price of anarchy revisited). *Let* $\mathcal{S}$ *denote the class of cost mappings that are monotone gradient mappings and for which the system cost* S *is convex. Then*

$$r(\Delta) \leqslant \sup_{\mathbf{0} \neq \mathbf{t} \in \mathcal{S}} \sigma(\mathbf{t})$$

holds.

Note that the above result is mainly of theoretical interest, as the computation of the steepness appears intractable for general asymmetric mappings. Moreover, these bounds are unlikely to be tight in practical situations.

2.10 Bibliographical notes

While Wardrop was first in providing a complete mathematical description of the user equilibrium conditions, the first to use the term *equilibrium* to describe the traffic pattern is perhaps the economist Knight (1924). The early

historical development of the equilibrium concept within the transportation field has been traced in Boyce (2004) and Patriksson (2006).

The presentation in Sections 2.1–2.8 follows that of the forthcoming book by Patriksson (2006). The modeling development differs from the traditional one (e.g., Aashtiani and Magnanti, 1981; Ferris and Pang, 1997; Facchinei and Pang, 2003a) in which the elastic demand Wardrop conditions are transformed into a nonlinear complementarity problem (NCP). The reason why we adopted the *mixed* complementarity approach is that it is the most natural, and that it possesses stronger properties: in the case of the NCP model, the equivalence with the Wardrop conditions requires that travel cost function be nonnegative everywhere – this is associated with having to impose the requirement that $\boldsymbol{\pi} \geqslant \mathbf{0}^{|C|}$ holds. In the MCP model, such a condition is not needed. For a general introduction to NCP and MCP problems, see Ferris and Pang (1997) and Facchinei and Pang (2003a).

The fixed demand reformulation and Proposition 1, presented in Section 2.4, stem from Patriksson (2006), and extend the work of Gartner (1980) on problems with a separable costs structure. Sheffi (1985, Section 6.3) remarks that the selection of a good upper bound $\bar{\mathbf{g}}$ on the demand function is crucial to the performance of the fixed-demand algorithm, and proposes means to iteratively update it during the computational process, when a linearization (Frank–Wolfe) strategy is adopted.

The Flow Decomposition Theorem 2 can be found, for example, in Busacker and Saati (1965, Theorems 3.7 and 7.2), Rockafellar (1984, Sections 4.B and 4.J), or Ahuja et al. (1993, Theorem 3.5). The linear program (13) which can be used to decompose a link flow into the sum of a strongly acyclic flow and a circulation flow is due to Gallager (1977). Proposition 3 in Section 2.5 stems from Hagstrom and Tseng (1998); its extension to the case of elastic demands is immediate and is found in Patriksson (2006).

The existence results of Section 2.7 (taken from Patriksson (2006)) are new. This is surprising, given that variational inequality models of traffic equilibrium problems have been around since the late 1970s (e.g., Smith, 1979; Dafermos, 1980) and the optimization models in Section 2.6, corresponding to problems with separable costs and demands, have been around since the 1950s (Beckmann et al., 1956). The improvements have been made possible for two reasons; for the case of the link–route representation, we have already remarked that the cost function need not be nonnegative. For the link–node representation, we have been able to utilize the recent results on the acyclic nature of equilibria, stated in Proposition 3. Cost additivity was assumed in our presentation for simplicity only.

Some of the uniqueness results (taken from Patriksson (2006)) are likewise new, in particular regarding costs and demands. The classical result is that $\mathbf{t}$ monotone, $-\mathbf{g}$ strictly monotone $\Longrightarrow \boldsymbol{\pi}$, $\mathbf{d}$ unique (e.g., Florian, 1979; Aashtiani and Magnanti, 1981). That the cost mapping of a variational inequality is constant over the solution set whenever the mapping is monotone$^+$

(see Crouzeix et al., 2000) is used to good effect, as well as the result that "**f** monotone$^+$" coincides with "**f** monotone" whenever **f** is a gradient.

The content of Section 2.8 is taken from Patriksson (2006). It is clear that under continuity and some form of monotonicity of the travel cost function, all definitions collapse to the traditional one. Proposition 8 is collected partly from the following sources:

(a) For the first result, see Smith (1984, Theorem 1). See also Dafermos (1971, Theorem 2.6; 1982) (the latter for the elastic demand case). For the second result, see Heydecker (1986, Theorem 7). For the third result, see Dafermos (1971, Theorem 2.6) and Heydecker (1986).

(b) That equilibrated flows are Wardrop equilibria is established in Heydecker (1986, Theorem 5). Our result improves upon it by reducing the requirements to pseudomonotonicity.

(c) See Bernstein and Smith (1994, Theorem 2.1).

(d) See de Palma and Nesterov (1998, Theorem 4(i)); however, their result is weaker in that monotonicity is assumed.

In Section 2.9, the worst-case behavior of Wardrop equilibria, with respect to a system-optimal flow pattern, has been analyzed in Roughgarden and Tardos (2002), Chau and Sim (2003), Correa et al. (2004). We remark that the result that user and system optimal solutions coincide for the exponent $\mu = 0$ was established first by Jorgensen (1963), while the result for $\mu \geqslant 1$ is due to Dafermos and Sparrow (1969, Equation (1.34)); see also Bennett (1993).

3 Variations

3.1 Multi-mode and multi-attribute models

A model of traffic equilibrium where user behavior is captured by a single function, either route-based or link-based, is a mathematical idealization of real life. It fails to account for travel time (mis-)perceptions and for the fact that delay is but one attribute of travel, together with cost, driving stress, landscape, number of stops, vehicle type, etc. The first issue, mis-perception, is best dealt with by using stochastic models (logit, mixed logit, probit) while the second issue yields multi-mode or multi-attribute[12] models. In multi-mode models, distinct cost functions $\mathbf{t}_m$ are assigned to distinct modes $m \in \mathcal{M}$. One possible functional form is

$$\mathbf{t}_m\big(\mathbf{v}^1, \ldots, \mathbf{v}^{|\mathcal{M}|}\big) = \mathbf{t}_m\left(\sum_{i=1}^{|\mathcal{M}|} \alpha_i \mathbf{v}^i\right), \tag{25}$$

[12] We refrain from using the term 'multi-criterion' that usually refers to vector optimization or vector equilibrium problems.

where $\mathbf{v}^m$ denotes the flow of mode m and α_i may be interpreted as the car-equivalent associated with mode m. An equilibrium is then characterized as a solution to the variational inequality

$$-\mathbf{t}_m\big(\mathbf{v}^1,\dots,\mathbf{v}^{|\mathcal{M}|}\big) \in N_{F_m}\big(\mathbf{v}^1,\dots,\mathbf{v}^{|\mathcal{M}|}\big), \quad m \in \mathcal{M}, \tag{26}$$

where F_m denotes the set of demand-feasible flows for mode m. The theoretical drawbacks of such a functional form have been pointed out by some researchers, who suggested more complex but better behaved functional forms. Multi-mode models can easily be converted to the standard model by assigning to each mode its own copy of the transportation network, with a cost function that becomes nonseparable, and most likely asymmetric.

Closely related to multi-mode models are multi-attribute (or multiclass) models, where the disutility (generalized cost) of each user is a function of route attributes, and the perception of disutility associated with each attribute varies across the population. To fix ideas, let us consider a model with two attributes, namely travel time $\mathbf{t}$ and travel cost $\mathbf{f}$, the latter including both out-of-pocket cost (tolls, for example) and variable costs such as petrol or maintenance. Let us adopt monetary cost as numéraire, define α_g as the value of one time unit (VOT) associated with population group $g \in \mathcal{G}$, and make the hypothesis that the disutility of each class g assumes the linear form

$$u_g(\mathbf{v}) = \alpha_g \mathbf{t}(\mathbf{v}) + \mathbf{f}(\mathbf{v}), \tag{27}$$

where $\mathbf{v} = (\mathbf{v}^1,\dots,\mathbf{v}^{|\mathcal{G}|})$. Let us denote by d_k^g the set of demand-feasible flow for population group g and origin–destination pair k, and by F_g the corresponding set of demand-feasible link flows. Due to the separability of the feasible sets, the equilibrium is then characterized by the variational inequalities

$$-u_g\big(\mathbf{v}^1,\dots,\mathbf{v}^{|\mathcal{G}|}\big) \in N_{F_g}\big(\mathbf{v}^g\big), \quad g \in \mathcal{G}. \tag{28}$$

Although the above model is subsumed, from the mathematical point of view, by the variational inequality (26), its analysis yields insight into models involving distinct classes of vehicles or customers, and is well worth investigating for its own sake. As an introductory example, let us consider a two-group, one-link example, with cost mapping

$$\mathbf{u}\big(v^1, v^2\big) = \begin{pmatrix} \alpha_1 t(v^1+v^2) + f(v^1+v^2) \\ \alpha_2 t(v^1+v^2) + f(v^1+v^2) \end{pmatrix} \tag{29}$$

and with Jacobian matrix

$$\nabla\mathbf{u}\big(v^1, v^2\big) = \begin{pmatrix} \alpha_1 t'(v^1+v^2) + f'(v^1+v^2) & \alpha_1 t'(v^1+v^2) + f'(v^1+v^2) \\ \alpha_2 t'(v^1+v^2) + f'(v^1+v^2) & \alpha_2 t'(v^1+v^2) + f'(v^1+v^2) \end{pmatrix}. \tag{30}$$

It is instructive to check the monotonicity of the mapping $\mathbf{u}$. To this aim, we compute the determinant of the symmetrized Jacobian matrix:

$$\begin{aligned}
&\det\bigl(\nabla\mathbf{u}+\nabla\mathbf{u}^{\top}\bigr)\\
&\quad=\begin{vmatrix} 2\alpha_1 t'(v^1+v^2)+2f'(v^1+v^2) & (\alpha_1+\alpha_2)t'(v^1+v^2)+2f'(v^1+v^2)\\ (\alpha_1+\alpha_2)t'(v^1+v^2)+2f'(v^1+v^2) & 2\alpha_2 t'(v^1+v^2)+2f'(v^1+v^2)\end{vmatrix}\\
&\quad=\bigl[4\alpha_1\alpha_2-(\alpha_1+\alpha_2)^2\bigr]\bigl(t'\bigl(v^1+v^2\bigr)\bigr)^2=-(\alpha_1-\alpha_2)^2\bigl(t'\bigl(v^1+v^2\bigr)\bigr)^2.
\end{aligned}$$

This calculation shows that $\mathbf{u}$ can only be monotone if either $\alpha_1=\alpha_2$ or the cost function t is constant. In the first case, the problem reduces to the single-group case while, if t is constant, congestion is absent. Now, since the feasible domain is separable by group, one may scale each group cost mapping by an arbitrary positive number, to derive an equivalent formulation of the equilibrium conditions. For instance, if the parameters α_1 and α_2 are not zero, we can divide the cost functions by their respective VOT parameter, yielding the equivalent cost mapping, which we still denote by $\mathbf{u}$:

$$\mathbf{u}\bigl(v^1,v^2\bigr)=\begin{pmatrix} t(v^1+v^2)+\alpha_1 f(v^1+v^2)\\ t(v^1+v^2)+\alpha_2 f(v^1+v^2)\end{pmatrix}, \tag{31}$$

and where α_g represents the inverse VOT of group g. Although both formulations are equivalent from the equilibrium viewpoint, their mathematical properties are quite different. Indeed, repeating the previous analysis, we obtain that $\mathbf{u}$ is monotone if f (rather than t previously) is constant. This will be satisfied if tolls and variable travel costs are flow-independent, a reasonable first-approximation assumption. For this reason we will, from now on, switch to *travel time* as numéraire. Note that the monotonicity results extend to the general multi-attribute case, i.e., a sufficient condition that the cost mapping $\mathbf{u}$ be monotone is that all cost functions, with the possible exception of travel time, be flow independent. We shall see, in a more general setting, that this condition is necessary as well.

A natural question that arises concerning the variational inequality involving the mapping $\mathbf{u}$ is the existence of a 'natural' optimization problem whose set of stationary (Karush–Kuhn–Tucker) points coincides with the set of multi-attribute equilibria. In other words: under what conditions is the mapping $\mathbf{u}$ a gradient mapping? Let $\mathcal{C}$ denote the index set of attributes and $\bar{\mathbf{v}}=\sum_{i\in\mathcal{C}}\mathbf{v}^i$ the total link flow vector. The answer to our question is that $\mathbf{f}$ be constant and $\mathbf{t}$ be a gradient mapping, and the relevant optimization program is simply to

$$\underset{\mathbf{v}}{\text{minimize}}\int_0^{\bar{\mathbf{v}}}\mathbf{t}(s)\,\mathrm{d}s+\sum_{i\in\mathcal{C}}\alpha_i\mathbf{f}^{\top}\mathbf{v}^i, \tag{32}$$

where $\mathcal{C}$ denotes the index set of criteria. Indeed, it is easy to check that the optimality conditions of (32) fulfill the multi-attribute equilibrium conditions. Furthermore, if the mapping $\mathbf{t}$ is monotone, the program (32) is convex. If $\mathbf{t}$ is

strictly monotone, the total flow equilibrium is unique, although group and origin–destination flows will not in general be unique.

We now focus our attention on a two-attribute model where the VOT (or inverse VOT) parameter is continuously distributed across the population, and described by a continuous probability density function h defined over the non-negative axis. For ease of notation, we assume that the density h is identical for all origin–destination pairs. For each origin–destination pair k, we then have that the demand density for the α-group[13] is equal to $h(\alpha)d_k$, and $F_\alpha = h(\alpha)F$ is the feasible set for the α-group. The variables of the equilibrium problem are regrouped into a link flow density vector $\mathbf{v}(\alpha) = \{v_l(\alpha)\}_{l\in\mathcal{L}}$ and, similar to the finite-dimensional case, we denote by $\bar{\mathbf{v}}$ the total link flow vector which, in this case, is defined by the integral

$$\bar{\mathbf{v}} = \int_0^\infty \mathbf{v}(\alpha)\,\mathrm{d}\alpha. \tag{33}$$

Replacing the index g by α yields the equilibrium conditions:

$$-\big(\mathbf{t}(\bar{\mathbf{v}}) + \alpha\mathbf{f}(\bar{\mathbf{v}})\big) \in N_{F_\alpha}\big(\mathbf{v}(\alpha)\big), \quad \alpha \in [0,\infty). \tag{34}$$

The above formulation involves an infinite collection of variational inequalities, one for each α-group. Upon introduction of the vectorial disutility function $\mathbf{u}(\bar{\mathbf{v}})$ defined as

$$\big[\mathbf{u}(\bar{\mathbf{v}})\big](\alpha) = \mathbf{t}(\bar{\mathbf{v}}) + \alpha\mathbf{f}(\bar{\mathbf{v}}), \tag{35}$$

the vector function $\mathbf{v} = \mathbf{v}(\alpha)_{\alpha\geqslant 0}$ and the set $F = \prod_\alpha F_\alpha$, the equilibrium conditions (34) can be aggregated into the infinite-dimensional variational inequality

$$-\mathbf{u}(\bar{\mathbf{v}}) \in N_F(\mathbf{v}). \tag{36}$$

While this may be considered elegant from the mathematical point of view, the approach has drawbacks. First, the proper setting for such formulation is the set of square-integrable functions. Since two functions that differ over a set of measure zero are equivalent, standard existence results cannot be invoked to prove the existence of a solution to the system (35) for *every* value of the parameter α. Second, this framework hides the network structure of the problem, which is essentially discrete.

Both these drawbacks can be remedied by considering finite-dimensional formulations. The remainder of the section is devoted to this topic, while algorithms will be discussed in Section 5.1. To ease the presentation, we assume from now on that F denotes the unit simplex $F = \{\bar{\mathbf{v}}_i \geqslant 0 \mid \sum_{i=1}^n \bar{\mathbf{v}}_i = 1\}$ and that $h(\alpha) = 0$ for all α larger than some finite threshold $\bar{\alpha}$. Since the original set F is a bounded polyhedron, the simplicial form of F can be achieved by expressing each of the points of F as a convex combination of its vertices.

[13] Although the terminology α-group is convenient, the notion actually relates to an infinitesimal user.

Alternatively, one may consider a traffic equilibrium problem involving one origin, one destination and a unit demand. For this reason, we will use, without any loss of generality, the term 'path' to denote an extreme point of the unit simplex.

Our finite-dimensional approach is based on the fixed point formulation of a variational inequality. In our setting, this takes the form:

$$v(\alpha) \in h(\alpha)F \cap \arg\operatorname*{minimum}_{\mathbf{w}(\alpha)\in F_\alpha} \big(\mathbf{t}(\bar{\mathbf{v}}) + \alpha\mathbf{f}(\bar{\mathbf{v}})\big)^\top \mathbf{w}(\alpha), \quad \alpha \in [0,\infty). \tag{37}$$

A key observation is that the right-hand side of (37) is equivalent to the set of linear programs

$$\operatorname*{minimize}_{\mathbf{w}\in F} \big(\mathbf{t}(\bar{\mathbf{v}}) + \alpha\mathbf{f}(\bar{\mathbf{v}})\big)^\top \mathbf{w}(\alpha), \quad \alpha \in [0,\infty). \tag{38}$$

This is nothing else than a parametric linear program, where the parameter α scans the nonnegative real axis. Because the number of extreme points of F, that is, combinations of shortest paths, is finite, this infinite-dimensional program can actually be solved in *finite* time by the parametric simplex algorithm. To this aim, let us sort the paths in decreasing order of their slopes. As the valuation of time decreases with α, the cost-conscious users will use high-index paths, while the time-conscious users will travel on low-index paths. The solution to the parametric linear program is characterized by a vector $\boldsymbol{\alpha} = (\alpha_0 = 0, \alpha_1, \ldots, \alpha_n = \bar{\alpha})$ of critical values, where users having an inverse VOT α belonging to the interval (α_{i-1}, α_i) are assigned to path i (see Figure 5). Several comments are in order:

(i) The flow on path i is given by the integral

$$\bar{\mathbf{w}}(\bar{\mathbf{v}}) = \int_{\alpha_{i-1}}^{\alpha_i} h(\alpha)\,\mathrm{d}\alpha. \tag{39}$$

If the total flow vector $T(\bar{\mathbf{v}}) = (T_i(\bar{\mathbf{v}}))_{i=1}^n$ agrees with $\bar{\mathbf{v}}$, then the solution to the parametric LP is an equilibrium for the bi-attribute problem. Since this solution only depends on $\bar{\mathbf{v}}$, it can be recovered from the solution to the finite-dimensional fixed-point problem $\bar{\mathbf{w}}(\bar{\mathbf{v}}) \in \bar{\mathbf{v}}$. Alternatively, a finite-dimensional fixed point formulation can be built around the vector of critical points $\boldsymbol{\alpha}$. This amounts to initializing with $\boldsymbol{\alpha}$ the sequence of evaluations

$$\text{total flow } \bar{\mathbf{v}} \rightarrow \text{ critical vector } \boldsymbol{\alpha} \rightarrow \text{ (through integration) } \bar{\mathbf{w}}.$$

(ii) The assignment of users whose inverse VOT is equal to one of the critical points $(\alpha_1, \ldots, \alpha_{n-1})$ is ambiguous. Generically, the number of such points is finite, i.e., of zero Lebesgue measure.

(iii) In the degenerate situation where two or more paths have identical $\mathbf{t}$ and $\mathbf{f}$ values over an interval (α_{i-1}, α_i), users may be assigned to any of these paths, a situation similar to that occurring in the standard, single-attribute model.

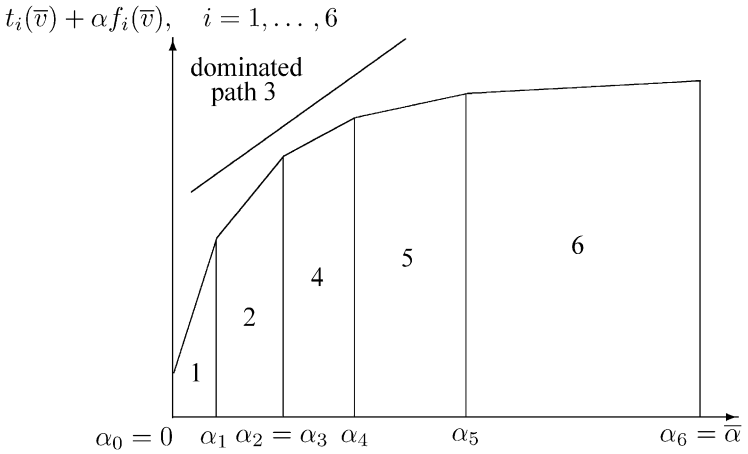


Fig. 5. Parametric path costs. Path 3 is dominated and therefore carries no flow.

(iv) If all slopes are distinct, the path assignment is unambiguous, with the exception of at most $n-1$ values of the parameter α, and the solution of the parametric LP is unique, almost everywhere. This regularity condition, which is easily achieved by perturbing the f-value of each path (or link), allows for a finite-dimensional variational inequality formulation and has beneficial algorithmic implications.

We now consider four important issues related to any fixed point or variational formulation, namely: existence and uniqueness of solutions, monotonicity and the gradient property. In the infinite-dimensional setting, existence of a solution can be proved using measure-theoretic arguments. In either finite-dimensional fixed point formulation, one can actually show that the fixed point mapping $\bar{\mathbf{w}}$ is upper semicontinuous whenever the ordering of path slopes remain constant, and existence follows from Kakutani's theorem. As in the discrete case, monotonicity holds if and only if $\mathbf{t}$ is monotone and $\mathbf{f}$ is constant. Similarly, the gradient property is satisfied if and only if $\mathbf{t}$ is a gradient mapping (not necessarily monotone) and, again, $\mathbf{f}$ is constant. The equivalent convex optimization problem is to

$$\underset{\bar{\mathbf{v}} \in F}{\text{minimize}} \int_{\mathbf{0}}^{\bar{\mathbf{v}}} \mathbf{t}(s)\, \mathrm{d}s + \sum_{i=1}^{n} \left[(f_i - f_{i+1}) \nu \left(\sum_{j=1}^{i} \bar{\mathbf{v}}_j \right) \right], \tag{40}$$

where $\nu = \int \phi^{-1}$ and ϕ denotes the cumulative distribution function ϕ of the density h.

The case for uniqueness deserves more attention. If $\mathbf{t}$ is constant and $\mathbf{f}$ is strictly monotone, the (total flow) solution $\bar{\mathbf{v}}$ to is unique. Moreover, if path slopes are distinct, the solution to the parametric LP (37) is unique, except for the critical points α_i. It follows that path flows are unique, a surprising result.

Indeed, the Jacobian matrix $\nabla \mathbf{u}(\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^{|\mathcal{G}|})$ of the disutility function $\mathbf{u}$ has rank one and is therefore 'less and less' definite as the number of user groups increases, making futile the standard argument for establishing uniqueness. Although the situation might be expected to worsen in the infinite-dimensional case, exactly the opposite occurs. This is reminiscent of stochastic assignment models, for which path uniqueness can be established under weak regularity assumptions.

3.2 Side constrained traffic equilibrium models

3.2.1 Introduction and equilibrium characterizations

Traffic flows are frequently subject to restrictions, physical or otherwise, that have not been taken into account in the basic model. For example, in the presence of a stationary traffic control system, flows might have to obey constraints of the form

$$s_k(\mathbf{v}) \leqslant 0, \quad k \in \mathcal{K}, \tag{41}$$

where, for mathematical convenience, we assume that the functions $s_k : \Re_+^{|\mathcal{L}|} \to \Re$, $k \in \mathcal{K}$, are convex and differentiable. Simple flow capacities on some links, describing the stationary effect of the traffic control, correspond to letting $\mathcal{K} := \overline{\mathcal{L}} \subseteq \mathcal{L}$, and

$$s_l(\mathbf{v}) := v_l - c_l, \quad l \in \overline{\mathcal{L}},$$

where $c_l > 0$ is the (stationary) link volume capacity. Side constraints can also be used to force equilibrium flows to comply with traffic management goals. For instance, system optimal link flow patterns can be enforced by setting c_l to the elements of a system optimal link flow vector.

Unfortunately, under the presence of joint constraints, the Cartesian product structure of the feasible sets of, for example, (15) and (16), are not satisfied any more. This has rather far reaching consequences on the equilibrium characterization of optimal flows. An equilibrium, in the sense of Wardrop, may no longer exist in terms of the original cost structure. If, however, one incorporates into the cost the Lagrange multipliers associated with the side constraints (41), then one can describe an equilibrium in terms of a generalized cost which, under some circumstances, can be given a natural interpretation. We introduce multipliers $\beta_k \geqslant 0$ for each constraint in (41) and append to (2) the additional condition

$$\mathbf{0}^{|\mathcal{K}|} \leqslant \boldsymbol{\beta} \perp -\mathbf{s}(\mathbf{v}) \geqslant \mathbf{0}^{|\mathcal{K}|}. \tag{42}$$

Letting $S := \{\mathbf{v} \in \Re^{|\mathcal{L}|} \mid s_k(\mathbf{v}) \leqslant 0, k \in \mathcal{K}\}$ the side constrained inelastic demand model becomes

$$-\mathbf{t}(\mathbf{v}) \in N_{F \cap S}(\mathbf{v}) = N_F(\mathbf{v}) + N_S(\mathbf{v}) = N_F(\mathbf{v}) + \boldsymbol{\nu}, \quad \boldsymbol{\nu} \in N_S(\mathbf{v});$$

if S satisfies a constraint qualification, for instance if each function s_k is affine, then for some vector $\boldsymbol{\beta}$ satisfying (42), we have that, at equilibrium,

$$-\left[\mathbf{t}(\mathbf{v}) + \sum_{k\in\mathcal{K}} \beta_k \nabla s_k(\mathbf{v})\right] \in N_F(\mathbf{v})$$

holds. Hence, for any optimal vector of Lagrange multipliers, $\mathbf{v}$ is a Wardrop equilibrium volume in terms of the generalized link costs $\bar{\mathbf{t}}(\mathbf{v}) := \mathbf{t}(\mathbf{v}) + \sum_{k\in\mathcal{K}} \beta_k \nabla s_k(\mathbf{v})$. In the case of link capacities, this reduces to $\bar{\mathbf{t}}(\mathbf{v}) := \mathbf{t}(\mathbf{v}) + \boldsymbol{\beta}$, where the elements of $\boldsymbol{\beta}$ for $l \notin \overline{\mathcal{L}}$ are fixed to 0. This expression may be interpreted as a generalized delay, where $\boldsymbol{\beta}$ is a vector of *queueing* delays at intersections. While this interpretation may be valid, it is important to note that the vector $\boldsymbol{\beta}$ of Lagrange multipliers is seldom unique, even in this simple case. This implies that link queueing delays are not unique, even if link flows $\mathbf{v}$ are. The following example illustrates how this situation can be utilized to improve traffic conditions.

We consider the directed graph illustrated in Figure 6, that involves two OD pairs, $(1, 5)$ and $(2, 5)$, with a demand of 2 and 3, respectively. Travel cost functions and link capacities are given in Table 1. The solution to the traffic equilibrium problem (that is, in the absence of upper bounds) is $\mathbf{v} = (2.00, 2.09, 0.91, 1.27, 2.82, 2.18)^\top$, while the solution to the capacitated model is $\mathbf{v}^* = (2, 2, 1, 2, 2, 3)^\top$.

Applying an augmented Lagrangian algorithm (see further Section 5.2) to this problem produces the multiplier vector $\boldsymbol{\beta} := (0, 0.219, 0.219, 0, 8.0, 0)^\top$. This is, however, far from the only multiplier vector in this problem. We first note that the set of multiplier vectors $\boldsymbol{\beta}$ for this problem satisfies

$$\begin{aligned}
&\beta_1 + \beta_4 + \beta_6 - \pi_{15} \geqslant -18,\\
&\beta_1 + \beta_5 - \pi_{15} \geqslant -10,\\
&\beta_2 + \beta_4 + \beta_6 - \pi_{25} \geqslant -24,\\
&\beta_2 + \beta_5 - \pi_{25} \geqslant -16,\\
&\beta_3 + \beta_6 - \pi_{25} \geqslant -24,\\
&2\beta_1 + 2\beta_2 + \beta_3 + 2\beta_4 + 2\beta_5 + 3\beta_6 - 2\pi_{15} - 3\pi_{25} = -92,\\
&\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \geqslant 0,\\
&\beta_6 = 0.
\end{aligned}$$

Clearly, this set is not a singleton, and is even unbounded. We see that link 6 is overcapacitated at $\mathbf{v}^*$, whence its multiplier must be zero by complementarity, while the remaining link multipliers are nonnegative, by definition. If the model is valid, it is then interesting to note that we may interpret the multipliers β_l as link tolls. These ensure that the uncapacitated traffic equilibrium problem with link cost $\mathbf{t}(\cdot) + \boldsymbol{\beta}$ has the same solution $\mathbf{v}^*$ as the capacitated model with the original link travel costs. Actually, the set of multipliers shown

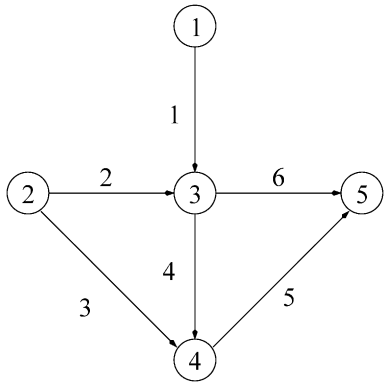


Fig. 6. A capacitated traffic network.

Table 1.
Network data

	Link	$t_l(v_l)$	c_l
1:	(1, 3)	$t_1(v_1) = v_1$	2
2:	(2, 3)	$t_2(v_2) = 4v_2$	2
3:	(2, 4)	$t_3(v_3) = 12v_3$	1
4:	(3, 4)	$t_4(v_4) = 2v_4$	2
5:	(3, 5)	$t_5(v_5) = 4v_5$	2
6:	(4, 5)	$t_6(v_6) = 4v_6$	4

above is only a subset of interesting choices. It may even turn out to be beneficial to impose *negative* tolls, or to set tolls on links that are *not* saturated, in order to achieve some preset management target. This possibility brings forward an alternative perspective on how to induce a favored equilibrium solution.

Suppose that we wish to impose upon the users link tolls (be they positive or negative) such that a link flow $\mathbf{v}^* \in \widehat{F}$ is in equilibrium. Then, if one interprets the side constraints as an enforcement of the equality $\mathbf{v} = \mathbf{v}^*$ (that is, $s_l(\mathbf{v}) = v_l - v_l^*$, $l \in \mathcal{L}$, and $s_{|\mathcal{L}|+l}(\mathbf{v}) = v_l^* - v_l$, $l \in \mathcal{L}$), one observes that the set of link tolls corresponds to the polyhedron defined by the solutions in $\boldsymbol{\beta}$ of the linear system

$$\boldsymbol{\Lambda}^\top \big(\mathbf{t}(\mathbf{v}^*) + \boldsymbol{\beta}\big) \geqslant \boldsymbol{\Gamma}\boldsymbol{\pi}, \tag{43a}$$

$$\big(\mathbf{t}(\mathbf{v}^*) + \boldsymbol{\beta}\big)^\top \mathbf{v}^* = \mathbf{d}^\top \boldsymbol{\pi}. \tag{43b}$$

In the above example, this set corresponds to the removal of the last two conditions, i.e., nonnegativity and complementarity.

Using the previous polyhedron as the feasible set of a mathematical program, we may, for example, devise a minimum-revenue toll by minimizing the

revenue function $\boldsymbol{\beta} \mapsto \boldsymbol{\beta}^\top \mathbf{v}^*$, whose minimal value 16 is achieved for the toll vector $\boldsymbol{\beta}^* = (0, 0, 0, 0, 8, 0)^\top$ (compare with the value 16.657 obtained from the augmented Lagrangian solution mentioned above) and $\boldsymbol{\pi} = (18, 24)^\top$. Obviously, there does not exist a maximum-revenue toll schedule, since the revenue can be made arbitrarily large along the line $\boldsymbol{\beta} = (0, \gamma, \gamma, 0, 8, 0)^\top$, $\gamma \geqslant 0$. (In the case of the set (43), we can let $\gamma \to -\infty$ to prove that no minimum-revenue toll exists.) Such toll problem will be investigated in more detail in Section 6.2.

In the elastic demand case, the toll polyhedron that induces given link flow and demand patterns $(\mathbf{d}^*, \mathbf{v}^*)$ is

$$\boldsymbol{\Lambda}^\top \big(\mathbf{t}(\mathbf{v}^*) + \boldsymbol{\beta}\big) \geqslant \boldsymbol{\Gamma}\boldsymbol{\pi}, \tag{44a}$$

$$\big(\mathbf{t}(\mathbf{v}^*) + \boldsymbol{\beta}\big)^\top \mathbf{v}^* = \boldsymbol{\xi}(\mathbf{d}^*)^\top \mathbf{d}^*, \tag{44b}$$

that is, the immediate extension to the system (43). We note that the value of $\boldsymbol{\beta}^\top \mathbf{v}^*$ is *constant* over this set, i.e., the toll revenue is constant! This apparent paradox can be explained by the fact that network users might *not* travel when the travel cost becomes too large.

3.2.2 Wardrop-like principles

According to the above, side constrained equilibria satisfy the Wardrop conditions in terms of *generalized* travel costs, but they will, in general, not satisfy any similar conditions in terms of *actual* travel costs. One can therefore, in general, not relate the actual travel costs of the unused routes to those of the used ones; for example, the least costly route in an OD pair may be unused because its generalized cost is too high. This deficiency is due to the fact that the side constrained problem does, in contrast to the standard model, lack a Cartesian product structure.[14]

Wardrop-like principles in terms of actual travel costs may however be established if the following assumption is fulfilled at the solution to the side constrained model.

Assumption 12 (Nondecreasing side constraint functions). At the flow $\mathbf{v} \in \widehat{F}$,

$$\frac{\partial s_k(\mathbf{v})}{\partial v_l} \geqslant 0, \quad l \in \mathcal{L},\ k \in \mathcal{K},$$

holds.

If this assumption holds for any flow $\mathbf{v} \in \widehat{F}$, then a flow increase along one or more links of the network can never result in any side constraint becoming 'looser'. Conversely, this assumption holds whenever the side constraints correspond to general capacity restrictions, that is, when they represent upper

[14] For capacitated networks, an alternative approach is analyzed in Section 3.3.

bounds on traffic volumes on certain links or routes, or within an area of the traffic system. In the sequel, we shall use the notions of links and routes that are *unsaturated* with respect to the side constraints.

Definition 13 (Unsaturated link and route). A link $l \in \mathcal{L}$ is unsaturated at the flow $\mathbf{v} \in \widehat{F}$ if for all $k \in \mathcal{K}$,

$$\frac{\partial s_k(\mathbf{v})}{\partial v_l} > 0 \quad \Longrightarrow \quad s_k(\mathbf{v}) < 0.$$

A route $r \in \mathcal{R}$ is unsaturated at the flow $\mathbf{v} \in \widehat{F}$ if all links $l \in \mathcal{L}$ on route r are unsaturated.

A route is clearly *saturated* at the flow $\mathbf{v} \in \widehat{F}$ if $\partial s_k(\mathbf{v})/\partial v_l > 0$ holds for some link l on the route and some $k \in \mathcal{K}$ such that $s_k(\mathbf{v}) = 0$.

Theorem 14 (Wardrop-type principles). *Suppose that* $(\mathbf{h}^*, \mathbf{v}^*)$ *solves the side constrained model and that Assumption* 12 *holds at* $\mathbf{v}^*$. *Then, the following conclusions hold for any OD pair* $(p, q) \in \mathcal{C}$.

(a) *The routes utilized in the OD pair have equal and minimal generalized route costs.*

(b) *Assume, without loss of generality, that the first* ℓ *routes are utilized in the OD pair and that m of these are unsaturated. Then, the routes may be ordered so that*

$$c_1 = c_2 = \cdots = c_m \geqslant c_{m+1} \geqslant c_{m+2} \geqslant \cdots \geqslant c_\ell.$$

(c) *For any pair of routes* $r, s \in \mathcal{R}_{pq}$,

$$\left.\begin{array}{l} \textit{route } r \textit{ is unsaturated} \\ c_s > c_r \end{array}\right\} \quad \Longrightarrow \quad h_s^* = 0.$$

(d) *For any pair of routes* $r, s \in \mathcal{R}_{pq}$,

$$\left.\begin{array}{l} \textit{route } r \textit{ is utilized} \\ c_s < c_r \end{array}\right\} \quad \Longrightarrow \quad \textit{route } s \textit{ is saturated}.$$

The (simple) proof of the above results is based upon the equilibrium characterization of a side constrained equilibrium.

If the implication in either of the results (c) and (d) were not fulfilled for some pair of routes, then some traveler might shift to a less costly and unsaturated alternative route; hence, these results are quite natural. As touched upon above, the OD routes that are unused in a solution to the side constrained model are not necessarily more costly (in actual travel cost) than those used in the OD pair; this is implied by the result (d) since a route may be saturated at zero flow.

3.3 Strategic models

The standard traffic equilibrium models assume that path selection, performed at the origin nodes, does not change en-route. This assumption is reasonable as a first approximation, but is not entirely compatible with user behavior. In this section, we outline an approach that allows users to take online decisions, based on current traffic conditions. While such an approach is common place in transit systems, where users may need to transfer one or more times to reach their destination, its application to congested traffic assignment is recent.

Let us first consider a transit system where, at each stop served by several transit lines, waiting customers must make strategic choices of the form: Should one board the incoming vehicle, or wait for another, more attractive (quicker) transit line? Such a decision must balance the travel time of the incoming vehicle against travel times *plus* waiting times of vehicles yet to reach the boarding station. Let us assume that the transit stop is served by two lines with respective frequencies ϕ_1, ϕ_2, and travel times to destination $t_1 < t_2$. Assuming that the arrival process is random and memoryless (Poisson process), and that users are rational and risk-neutral, they will naturally board a vehicle of line 1 if it shows up first, and a vehicle of line 2 if the expected waiting time of line 1, i.e., $1/\phi_1$, exceeds the difference $t_2 - t_1$, i.e.,

$$t_2 \leqslant \frac{1}{\phi_1 + t_1}.$$

This reasoning can be generalized to n common bus lines indexed in decreasing order of their travel times to destination. The optimal strategy is characterized by a threshold index $\bar{l}$, below which all transit lines are attractive, that satisfies the relation

$$\bar{l} \in \arg\operatorname*{minimum}_{1 \leqslant l \leqslant n} \frac{1}{\sum_{i=1}^{l} \phi_i} \left(1 + \sum_{i=1}^{l} \phi_i t_i \right). \tag{45}$$

The key feature of this approach is that, while strategies are *deterministic* objects, the route traveled from day to day by users is *stochastic*. The applicability of the method relies on the existence of efficient algorithms for computing optimal strategies. Such algorithms, akin to shortest path methods, are able to address large-scale problems.

In private transportation, the situation is different, since randomness is associated with demand (users) rather than supply (vehicles). Defining equilibrium meaningfully in this context represents a nontrivial task. To gain some insight into the situation, let us consider the network illustrated in Figure 7, where each link is endowed with a cost (shown next to the corresponding link) and, possibly, a capacity (bracketed number). Paths from origin node 1 to destination node 5 are listed, together with their features, in Table 2.

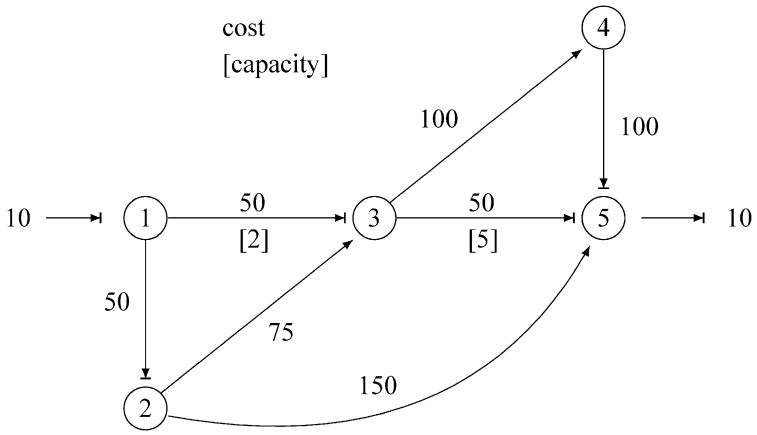


Fig. 7. A capacitated network.

Table 2.
Network paths for a small example

Path index	Path	Cost	Capacity
1	1–3–5	100	2
2	1–2–3–5	175	5
3	1–2–5	200	∞
4	1–3–4–5	250	2
5	1–2–3–4–5	325	∞

Clearly, no Wardrop equilibrium is compatible with this data, since the shortest path 1–3–5 cannot accommodate all demand of 10 units. Quite naturally, one could settle for an equilibrium satisfying the following Wardrop-like principle:

> At equilibrium, the cost of a path with a positive residual capacity is larger or equal to the cost of any path carrying positive flow.

In our small example, any path flow vector of the form

$$\mathbf{x} = (5 - \delta, \delta, 5, 0, 0)^\top$$

with $\delta \in [3, 5]$ satisfies the above principle. However, whenever δ is strictly greater than 3 and the users of the second path are endowed with the gift of prescience, it is tempting for them to switch to the less costly first path on which some capacity has been set free by *themselves*. In this sense, the equilibrium associated with the value $\delta = 3$ is more natural. Indeed this solution, associated with a queueing delay of 75 on link (1, 3) and 25 on link (3, 5), corresponds to the equilibrium notion of Section 3.2. Notwithstanding the queueing delay, it is a system-optimal solution as well.

Table 3.
A set of strategies for the small example

Node	1	2	3	4	5
s_1	[3, 2]	[3]	[5, 4]	[5]	[ ]
s_2	[3, 2]	[5]	[5, 4]	[5]	[ ]
s_3	[2]	[3]	[5, 4]	[5]	[ ]
s_4	[2]	[5]	[ ]	[ ]	[ ]
s_5	[3, 2]	[5, 3]	[5, 4]	[5]	[ ]
s_6	[3]	[ ]	[5, 4]	[5]	[ ]
s_7	[3, 2]	[3]	[4, 5]	[5]	[ ]

Let us now consider the strategic approach where a subset of strategies, represented as vectors whose elements consist of an ordered list of outgoing nodes, is provided in Table 3. For instance, when leaving its origin node 1, a user adopting strategy s_1 will prefer going to node 3 (first node in the ordered list) over the alternative choice, that is, going to node 2, provided that link (1, 3) is not saturated yet. A similar situation occurs at node 3, where the outgoing link (3, 5) is preferred over (3, 4). At nodes 2 and 4, however, the strategic choice shrinks to a singleton, due to the fact that the respective preferred outgoing links are uncapacitated. It follows that a user adopting strategy s_1 could end up traveling on path 1, 2, 4, or 5, depending on the availability of the capacitated links (1, 3) and (3, 5). The strategies that, such as s_4, avoid capacitated links, correspond to ordinary paths in the network.

Whenever the number v_a of users that wish to access link l exceeds its capacity u_l, the probability p of accessing the link is set to $p = u_l/v_l$. This is equivalent to assuming that users are independently and uniformly distributed at the tail node of link l, in a dimensionless queue. These *access probabilities* (or *access proportions*) allow us to compute the expected cost of strategies. For instance, if all 10 users adopt strategy s_1, the probability of accessing link (1, 3) is equal to 2/10. The users, whether they access link (1, 3) or not, clash again at node 3, where the access probability of link (3, 5) is 5/10. These numbers, which are required to compute the access probabilities associated with the paths of the network, are shown in Table 4.

The expected value of each user's *expected* delay is then equal to the sum of the path costs, weighted by the respective path access probabilities, i.e.,

$$\left(\frac{1}{10} \times 100\right) + \left(\frac{4}{10} \times 175\right) + \left(\frac{0}{10} \times 200\right) + \left(\frac{1}{10} \times 250\right) + \left(\frac{4}{10} \times 325\right) = 235.$$

A *strategic equilibrium* is reached when all users are assigned to strategies of minimal expected delays. This condition is clearly violated if all commuters are assigned to strategy s_1, since strategy s_4, which corresponds to the unca-

Table 4.
Path access probabilities for the small example

Path	Access probability	Cost
1–3–5	$\frac{2}{10} \times \frac{5}{10} = \frac{1}{10}$	100
1–2–3–5	$\frac{8}{10} \times \frac{5}{10} = \frac{4}{10}$	175
1–2–5	0	200
1–3–4–5	$\frac{2}{10} \times \frac{5}{10} = \frac{1}{10}$	250
1–2–3–4–5	$\frac{8}{10} \times \frac{5}{10} = \frac{4}{10}$	325

Table 5.
Equilibrium strategic path flows

Path	Flow from s_1	Flow from s_2	Cost
1–3–5	$\frac{5}{6}$	$\frac{5}{6}$	100
1–2–3–5	$\frac{20}{6}$	0	175
1–2–5	0	4	200
1–3–4–5	$\frac{1}{6}$	$\frac{1}{6}$	250
1–2–3–4–5	$\frac{4}{6}$	0	325

pacitated path 1–2–5, is available at cost $200 < 235$. It is not too difficult to verify that the unique equilibrium corresponds to the assignment of 5 users to strategy s_1 and of the remaining 5 to strategy s_2; the resulting strategic flows and path flows are shown in Table 5. One checks that the expected delay of each strategy is equal to 185, which is less than the expected delay of unused strategies, thus fulfilling the equilibrium conditions.

In general, a vector $\mathbf{x} = \{\mathbf{x}_k\}_{k \in \mathcal{C}}$ of *strategic flows* (one for each commodity or OD pair $k \in \mathcal{C}$) is a strategic equilibrium if and only if it is demand-feasible ($\mathbf{x}_k$ belongs to the set of feasible strategic flows X_k for every commodity $k \in \mathcal{C}$) and satisfies the variational inequality

$$-\mathbf{c}(\mathbf{x}) \in N_X(\mathbf{x}), \tag{46}$$

where X denotes the Cartesian product of the sets X_k. For a given strategy s, the component $\mathbf{c}^s(\mathbf{x})$ of $\mathbf{c}(\mathbf{x})$ represents the expected delay associated with strategy s and the strategic vector $\mathbf{x}$. Not that the evaluation of $\mathbf{c}$ requires the knowledge of total link flows. Unfortunately, in contrast with the standard model where this information is readily available from path flows (via the link–route incidence matrix), the situation is different in the strategic model, as the dependence of the link-hyperpath matrix on strategies complicates matters significantly. In particular, the task of determining strategic flows that are compatible with a given link flow vector is an algorithmic challenge by itself, and most likely 'intractable'.

In our introductory example, the link access probabilities were easy to compute. However, the situation becomes complex when the forward star of a node involves more than two links, and when strategies active at that node have different priority orders. In this situation, should one assume that users strictly obey a FIFO (First-In-First-Out) rule, or does one allow a user to jump the queue if its preferred choice is available, as in banks with tellers dedicated to specific services? The distinction is important, since it impacts the assignment of users to outgoing links. In the first case, users are assigned to their preferred node according to their position in the queue. In the second case, users are assigned *simultaneously* to their preferred node, until some residual capacity becomes zero. Throughout the process, users that are denied their current preferred link keep a priority compatible with their arrival instant. Both situations are illustrated in Figure 8 where each square, for ease of understanding, represents an atomic user whose preferred outgoing node is either A or B. The queue is 'virtual' (sometimes denoted 'vertical') in the sense that it occupies no physical space.

At node j, the loading of the flow onto the outgoing links is an iterative process. In the single queue case, it is initiated by constructing the set $\overline{K}$ of first choices. Next, one computes[15]

$$\eta = \underset{k\in\overline{K}}{\text{minimum}}\left\{\frac{\bar{u}_{jk}}{d_k}, 1\right\},$$

where d_k is the total demand for outgoing node k and $\bar{u}_{jk}$ is the residual capacity of link (j, k). If $\eta = 1$, all users access their preferred node k and the loading terminates trivially. Otherwise, let

$$\bar{k} \in \arg\underset{k\in\overline{K}}{\text{minimum}}\left\{\frac{\bar{u}_{jk}}{d_k}\right\}$$

denote the head node of the link (j, k) that gets saturated first. One then loads a fraction η of each demand to node k, updates demands, removes node k from the preference lists and repeats the process until demand is exhausted.

Let us consider a nontrivial example involving three outgoings links and two user groups of size 10 and 20 having respective preference lists $[k_1, k_2, k_3]$ and $[k_2, k_1, k_3]$. See Figure 9. We set the capacities of the three links to 8, 10, and 20. At the first iteration, link (j, k_2) gets saturated first; 5 flow units from strategy s_1 are assigned to k_1, and 10 units from s_2 to k_2. Next, one deletes node k_2 from the preference orders $E_j^{s_1}$ and $E_j^{s_2}$. At the second iteration, 15 units (5 from s_1 and 10 from s_2) select link (j, k_1), whose residual capacity is 3. A third iteration is required to terminate the process, as the 10 residual units are assigned to link (j, k_3). Note that a by-product of the loading process is the link access probabilities π_{jk} that allow to compute the strategic

[15] Here, the index k refers to a node, not an OD pair.

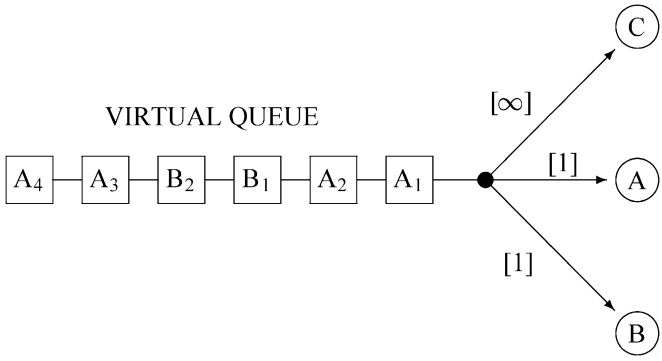


Fig. 8. Flow dispersion at a node. In the FIFO case, A_1 is assigned to A, A_2 is assigned to B and all other users are assigned to C. In the second case, A_1 is assigned to A while, in parallel, B_1 is assigned to B; the four remaining users are assigned to C.

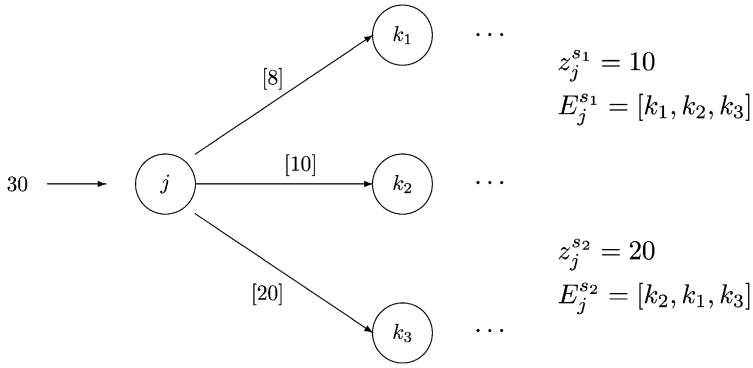


Fig. 9. Flow assignment at a node.

cost function **c**. The process is summarized in Table 6 while, based on the same data, the outcome of the 'parallel' assignment is summarized in Table 7.

Although there are considerable differences in the assignments resulting from the two rules, these agree in important situations, namely when at most two choices are available at each decision node. Other theoretical points worth mentioning are:

(i) The cost function **c** is continuous which, together with the compactness of the feasible set, implies that the set of equilibrium solutions is nonempty.

(ii) The loading procedure only makes sense if the network is acyclic, and nodes are processed in the relevant topological order. If this condition is not fulfilled, it is yet possible to perform the loading operation by iterating, à la Gauss–Seidel, with respect to the OD pairs. At a given

Table 6.
Outcome of single file loading at node j

	Link:	(j,k_1)	(j,k_2)	(j,k_3)
Iteration 1	Residual capacity:	8	10	20
	Flow:	$5(s_1)$	$10(s_2)$	0
Iteration 2	Residual capacity:	3	0	20
	Flow:	$5(s_1)+1(s_1)+2(s_2)$	$10(s_2)$	0
Iteration 3	Residual capacity:	0	0	20
	Flow:	$6(s_1)+2(s_2)$	$10(s_2)$	$4(s_1)+8(s_2)$
Probabilities:	$\pi^{s_1}_{jk_1}=\frac{6}{10}$, $\pi^{s_1}_{jk_3}=\frac{4}{10}$, $\pi^{s_2}_{jk_1}=\frac{2}{20}$, $\pi^{s_2}_{jk_2}=\frac{10}{20}$, $\pi^{s_2}_{jk_3}=\frac{8}{20}$			

Table 7.
Outcome of the parallel loading at node j

	Link:	(j,k_1)	(j,k_2)	(j,k_3)
Iteration 1	Residual capacity:	8	10	20
	Flow:	$8(s_1)$	$8(s_2)$	0
Iteration 2	Residual capacity:	0	2	20
	Flow:	$8(s_1)$	$8(s_2)+\frac{2}{7}(s_1)+\frac{12}{7}(s_2)$	0
Iteration 3	Residual capacity:	0	0	20
	Flow:	$8(s_1)$	$\frac{68}{7}(s_2)+\frac{2}{7}(s_1)$	$\frac{12}{7}(s_1)+\frac{72}{7}(s_2)$
Probabilities:	$\pi^{s_1}_{jk_1}=\frac{8}{10}$, $\pi^{s_1}_{jk_2}=\frac{2}{7\cdot 10}$, $\pi^{s_1}_{jk_3}=\frac{12}{7\cdot 10}$, $\pi^{s_2}_{jk_2}=\frac{68}{7\cdot 20}$, $\pi^{s_2}_{jk_3}=\frac{72}{7\cdot 20}$			

iteration, the access probabilities for all OD pairs, with the exception of the current one, are frozen at their previous values.

(iii) The cost mapping **c** may fail to be monotone under both queue disciplines.

(iv) In the single queue case, the issue of convexity of the equilibrium set is open.

From the practical side, static strategic models are of limited scope. However, they can be adapted to time-dependent networks and help in developing dynamic models that take into account the rational reaction of users to online information. In a dynamic context, models must take into account the natural temporal priorities; this leads to a more complex loading procedure involving a cost mapping **c** that may fail to be continuous. On the positive side, the mapping's upper semi-continuity ensures that the set of equilibria is nonempty, and the underlying time–space network is trivially acyclic.

3.4 Nonadditive route costs

In our previous discussions, we have assumed that route costs were additive, that is, $c_r(\mathbf{h}) = \sum_{l\in\mathcal{L}} \lambda_{lr} t_l(\mathbf{v})$ holds for all $r \in \mathcal{R}$ and consistent flows $(\mathbf{h}, \mathbf{v})$. This is not true, however, in several situations involving route-specific tolls or route-specific vehicle emissions. Cost additivity also fails in loss networks arising, among others, in telecommunication systems.

Let us reconsider the model with two attributes, time and money, introduced in Section 3.1. Empirical studies show that, in contrast with our previous assumption, the VOT parameter can be nonlinear. Consider, for example, that

$$c_r(\mathbf{h}) = f_r + \beta_r\left(\sum_{l\in\mathcal{L}} \lambda_{lr} t_l(\mathbf{v})\right), \tag{47}$$

where f_r is the monetary outlay and β_r is now a *function* of travel time, for route $r \in \mathcal{R}$. In this model, time is converted into money through the function β. It is also possible to work with time as the numéraire, thus considering instead a cost model of the form

$$c_r(\mathbf{h}) = \sum_{l\in\mathcal{L}} \lambda_{lr} t_l(\mathbf{v}) + \alpha_r(f_r), \tag{48}$$

where the scalar α_r converts money into time along route r. Interestingly, even if we set α_r to β_r^{-1}, the two models are *not* equivalent. From the analytic point of the view, the latter model has the advantage that there exists an equivalent optimization formulation in the case of separable link costs: provided that f_r is flow-independent, and that the value of ν_r does not vary within OD pair r, the partial derivative of the objective function

$$\sum_{l\in\mathcal{L}} \int_0^{v_l} t_l(s) + \sum_{k\in\mathcal{C}} \sum_{r\in\mathcal{R}_k} h_r \alpha_k(f_r)$$

with respect to h_r is exactly $c_r(\mathbf{h})$ given in (48), and Wardrop's conditions reduce to the optimality conditions of this optimization problem. We also see that the objective function is convex under the same conditions as for the additive model. Of course, these results hold because monetary costs are flow independent; observe the similarity between this result and those of Section 3.3.

It is clear that an algorithm in link-flow space based on first-order approximation cannot be immediately extended to this model; indeed the 'best' route cannot be determined through a standard shortest route calculation; see Section 5.4.

3.5 Bibliographical notes

References to the basic multi-mode model of Section 3.1 can be found in Florian (1977, 1979), Toint and Wynter (1996), Florian et al. (2002), Marcotte and Wynter (2004). The introduction of the bi-criterion model in the

transportation literature is due to Dial (1979), who also proposed a solution algorithm (Dial, 1996a, 1996b). Early on, Dafermos (1981) gave a necessary condition in order that the variational formulation of the problem be monotone. Similar results appear in Nagurney and Dong (2002). Marcotte (1998) proved that these conditions were necessary for both the monotonicity and integrability properties to hold. Equivalent finite-dimensional formulations of the problem have been proposed in Leurent (1993) and Marcotte (1998).

Section 3.2 introduces a modeling paradigm from the viewpoint of a control policy. This way of interpreting the model leads to interesting interpretations of the Lagrange multipliers, and the use of the 'side constraint' that $\mathbf{v} = \mathbf{v}^*$ must hold for some a priori set feasible flow $\mathbf{v}^*$ is the direct route to the use of link tolls for achieving a system optimum or any other desired flow, as presented, for example, in Section 6.2. Side constraints can emerge from other sources than controls, however, and Larsson and Patriksson (1994b, 1997, 1999) describe several such circumstances, and show how the Lagrange multiplier terms can be associated with link queues. That the Lagrange multipliers are not unique was established in Larsson and Patriksson (1998), whence interpretations must be used with great care, for example, in the context of equilibrium queues. Previously, equilibrium characterizations of link capacitated models have been provided in Jorgensen (1963) and Hearn (1980), and the interpretation of the multipliers β_l as stationary link queues can be found in Miller et al. (1975) and Payne and Thompson (1975). See also Ferrari (1995). The numerical example stems from Larsson and Patriksson (1998), as does the description of the toll polyhedron development. Related, and in many ways parallel, work exists for the special case where one wishes to achieve a *system-optimal* solution. In this field, the master's thesis by Bergendorff (1995) was followed by a series of work by Don Hearn et al. (e.g., Bergendorff et al., 1997; Hearn and Ramana, 1998).

The strategic approach outlined in Section 3.3 was motivated by the transit equilibrium model described in Nguyen et al. (2001), and is based on Marcotte et al. (2004). Additional information, including technical details, is available in Marcotte and Nguyen (1998), Hamdouch et al. (2004a, 2004b). Mathematical programming formulations of transit equilibrium problems, based on the notion of strategy (or hyperpath), were independently proposed in Nguyen and Pallottino (1988) and Spiess and Florian (1989).

The nonadditive models described in Section 3.4 are due to Gabriel and Bernstein (1997) (see also Chen and Bernstein, 2003), for the case (47) of converting money into time, and Larsson et al. (2002), for the case (48) of converting time into money. That the two models are not equivalent was observed in Larsson et al. (2002), and later established in more generality theoretically in Bernstein and Wynter (2000). The logit-based stochastic user equilibrium model is due to Fisk (1980). For further reading on stochastic traffic equilibrium models, see Sheffi (1985), Akamatsu (1996), Watling (1999), Cascetta (2001). The effects of modeling emission effects are discussed in Larsson et al. (2002). Recent models of loss networks in telecommunication can be found in

Altman et al. (2002). As is stated in the section, the model considered converts money into time, and models based on this transformation have better properties.

4 Solution algorithms: The basic problem

4.1 General guidelines

Finding a flow pattern satisfying Wardrop's user equilibrium conditions (2) amounts to solving a large-scale, network structured variational inequality problem. The computational challenge is threefold: How to exploit the underlying network structure of the problem efficiently? How to obtain fast local convergence? How to ensure global convergence?

Since the constraint set F, in the link–node representation, is the feasible set of an uncapacitated multicommodity flow problem, minimizing a linear function over F is a task that can be efficiently performed by finding the shortest route trees rooted at the origin nodes. At this level, the distinction between link, route or commodity flows becomes irrelevant, as the shortest route procedure provides information at all three levels of (dis)aggregation.

As far as convergence speed is concerned, algorithms combining the abovementioned linear subproblems with simple line searches only achieve sublinear convergence. To achieve linear or super-linear convergence in such a framework, one may have to solve a *quadratic* problem over the multicommodity network; unfortunately, such a task cannot be performed on large scale instances. Suppose, however, that we consider the link–route representation, and consider the set of routes carrying positive flow to be fixed. Then, solving the corresponding restricted problem by fast algorithms is indeed possible. Since, in practical applications, a relatively small set of routes is required to describe an equilibrium flow, this fact suggests that route (or, column) generation strategies are prime suspects for solving traffic equilibrium problems. It must be understood that the maximum size of the working set of routes will influence the behavior of the algorithm: the larger the maximum cardinality of this set, the better the convergence. However, a trade-off must be achieved between a good convergence rate on the one hand and, on the other hand, the size and the computational difficulty associated with the restricted problems. As a limiting case, the slow converging Frank–Wolfe method is obtained by limiting the size of the working set to two feasible points, one corresponding to the current iterate, and the second to the solution to the linear subproblem.

If $\mathbf{t}$ is not a gradient mapping, then one must be careful that an aggressive column dropping scheme will not result in a sequence of iterates that cycles between identical working sets. The third issue is more or less irrelevant if $\mathbf{t}$ is a gradient mapping, since the potential $\phi = \int \mathbf{t}$ then may be used to monitor descent toward a stationary (Karush–Kuhn–Tucker) point, that is, a user equilibrium. If $\mathbf{t}$ is *not* a gradient, then some monotonicity property (such as

pseudomonotonicity) is required in order to be able to implement a provably convergent method. In the following, we elaborate more on these issues for the basic traffic equilibrium model.

4.2 Model structures and the verification of equilibria

Some favorable problem structures are observed when looking at the models (15) and (16):

- *OD pair separability*. The feasibility of the demand and traffic volume in one OD pair does not affect that of another. The problems indeed are multicommodity flow problems, where there are no side constraints acting on more than one commodity, such as link flow capacity constraints. This property holds for more general cost and demand models as well.
- *Cost separability*. The objective function, as well as the link cost and demand functions, is separable in the link volume and demand variables. This property does not extend to more general cost and demand models.
- *Primal–dual relations*. In both the link–route and link–node representations, the vector $\boldsymbol{\pi}$ is a Lagrange multiplier vector, which further measures the least cost. This property holds for more general cost and demand models as well.

The link–route representation has a simpler constraint structure than the link–node one; the constraints in the former describe a simplex, while those of the latter describe flow conservation. The former, on the other hand, has an exponential number of variables, which must be enumerated iteratively, while the latter has a polynomial number of variables, and a richer constraint structure.

The very definition of a variational inequality provides a convenient criterion for verifying whether a vector is in equilibrium. Indeed, the inequality

$$\mathbf{f}(\mathbf{x})^{\top}(\mathbf{y}-\mathbf{x}) \geqslant 0, \quad \mathbf{x} \in X,$$

is satisfied if and only if $\mathbf{x}$ is optimal for the linear program

$$\underset{\mathbf{y}\in X}{\text{minimize}}\, \mathbf{f}(\mathbf{x})^{\top}\mathbf{y}.$$

If one keeps demand fixed, this amounts to checking whether the *gap function* (see also Appendix A)

$$\text{gap}(\mathbf{x}) = \underset{\mathbf{y}\in X}{\text{maximum}}\, \mathbf{f}(\mathbf{x})^{\top}(\mathbf{x}-\mathbf{y}) \tag{49}$$

is zero.[16] This LP problem defines a set of shortest route problems; an equilibrium is reached if and only if the total cost of transportation along the shortest

[16] The index k in the second equality relates to feasibility sets and subvectors for individual OD pairs.

routes in each OD pair equals the total travel cost of the current OD flow. The value of gap($\mathbf{h}$) is always nonnegative, and a flow vector $\mathbf{h}$ is an equilibrium if and only if gap($\mathbf{h}$) $= 0$. If gap($\mathbf{h}$) is positive, one has identified a set of attractive routes that are not currently used (at least not enough).

In the elastic demand case, it seems natural to consider the iteration

$$\mathbf{d} \mapsto (\mathbf{g} \circ \boldsymbol{\pi} \circ \mathbf{t} \circ \mathbf{v})(\mathbf{d}) \tag{50}$$

that takes the current demand and calculates a corresponding inelastic demand solution, from which the least route cost is calculated. The value of the demand function then yields the new demand estimate. A sufficient condition that this fixed-point iteration converge is that it be contractive, that is, for an equilibrium demand $\mathbf{d}^*$ and any feasible demand vector $\mathbf{d}$, there exists a scalar $\theta \in (0, 1)$ such that $\|\mathbf{g}(\boldsymbol{\pi}(\mathbf{t}(\mathbf{v}(\mathbf{d})))) - \mathbf{d}^*\| \leqslant \theta\|\mathbf{d} - \mathbf{d}^*\|$ holds. Given the complicated form of this mapping, whether the contraction condition is satisfied is difficult to analyze from the original data, and updating rules for flows and demands are instead based on approximations of one or more of these mappings, examples of which will be provided below.

4.3 Decomposition–coordination

The separability of the OD pairs (or, user classes, etc.) in the constraints suggests the use of a decomposition strategy, wherein flow of a given type is updated upon while keeping the others fixed; the remaining problem is then a single-commodity network flow problem. Among the advantages of such a scheme we mention first that the link cost function has stronger monotonicity properties when viewed on this smaller subspace than on the entire space of flows; for example, if $v_l \mapsto t_l(v_l)$ is strictly monotonically increasing, then it is (usually) nonstrictly monotone in the vector of OD link flows $\mathbf{w}_k$, $k \in \mathcal{C}$, but it is strictly monotonically increasing when keeping flows in all but one OD pair fixed. Second, several single-commodity network flow algorithms are available, among which are (potentially) superlinearly convergent ones. Third, algorithms of this type can be viewed as block versions of Gauss–Seidel (or Jacobi) algorithms, which have a sound convergence theory. They can however be slow when the number of blocks is large, since too much of the cost interaction among the commodities is lost. One can then take advantage of the problem structure by creating a small number of large blocks of variables such that, for example, all flow from one given origin is updated at any one time. Contrary to the use of decomposition over OD pairs or origins, a decomposition over links is less convenient, as costs are not always defined by separable functions.

A natural means to decompose the problem over OD pairs is to temporarily fix link costs, thereby removing the cost dependence between OD pairs. In the fixed demand problem, what is left is a number of shortest route problems, on which demand is assigned. In the elastic demand case, the demand to be assigned onto the shortest routes is given by the demand function evaluated

at the shortest route costs. In this way, we obtain the algorithms of Frank and Wolfe (1956) and Evans (1976), respectively.

4.3.1 Decomposition by Lagrangian relaxation

In order to take advantage of the link cost separability in the models (15) and (16), we must eliminate their dependence in the constraints by introducing some form of relaxation. One possibility is to introduce explicit multipliers, $\boldsymbol{\nu} \in \Re^{|\mathcal{L}|}$, for the constraints '$\mathbf{v} = \sum_{k\in\mathcal{C}} \mathbf{w}_k$'. With the (redundant) constraints $\mathbf{v} \in \Re_+^{|\mathcal{L}|}$ appended to the Lagrangian relaxed problem, we obtain a problem, equivalent to VIP($\mathbf{t}$, F_d), which amounts to finding ($\mathbf{v}$, $\mathbf{w}$, $\mathbf{d}$, $\boldsymbol{\pi}$, $\boldsymbol{\nu}$) such that

$$\mathbf{0}^{|\mathcal{L}|} \leqslant \mathbf{v} \perp \big(\mathbf{t}(\mathbf{v}) - \boldsymbol{\nu}\big) \geqslant \mathbf{0}^{|\mathcal{L}|}, \tag{51a}$$

$$\mathbf{E}\mathbf{w}_k = \mathbf{i}_k d_k, \quad k \in \mathcal{C}, \tag{51b}$$

$$\mathbf{0}^{|\mathcal{L}|} \leqslant \mathbf{w}_k \perp \big(\boldsymbol{\nu} - \mathbf{E}^\top \boldsymbol{\pi}_k\big) \geqslant \mathbf{0}^{|\mathcal{L}|}, \quad k \in \mathcal{C}, \tag{51c}$$

$$\mathbf{g}(\mathbf{d}) = \boldsymbol{\pi}, \tag{51d}$$

$$\mathbf{v} - \sum_{k\in\mathcal{C}} \mathbf{w}_k = \mathbf{0}^{|\mathcal{L}|} \tag{51e}$$

holds. Fixing $\bar{\boldsymbol{\nu}} \geqslant \mathbf{t}(\mathbf{0})$, and noting that (51e) is the dual optimality condition, the problem separates into one over $\mathbf{v}$ and $\mathbf{w}_k$, $k \in \mathcal{C}$. In the separable case, (51a) further decomposes into a set of single-variable problems with the objective $\int_0^{v_l} (t_l(s) - \bar{\nu}_l)\, \mathrm{d}s$, to be minimized over $\Re_+$, and with a solution given by the inverse of t_l evaluated at $\bar{\nu}_l$, whenever the latter is nonnegative (otherwise it is set to zero). In the general case, this is a nonlinear complementarity problem (NCP) in $\Re^{|\mathcal{L}|}$. The second problem, (51b)–(51d), amounts to finding, for each $k \in \mathcal{C}$, the shortest route based on the vector $\bar{\boldsymbol{\nu}}$ of link costs. The shortest route costs are used to calculate the demand through the demand function Demand is then distributed onto the shortest routes to form the vectors $\mathbf{w}_k$. A step is then taken along the direction $\boldsymbol{\nu}$, in order to produce a result that better fulfills the last constraint. Although being a primal–dual algorithm, the aggregate of the $\mathbf{w}_k$ is a feasible link flow; this is a consequence of the fact that the dualization is made with respect to the definitional constraint (51e).

Another way to decompose the problem is to relax the linear equality constraints defining the network, that is, (51b). Introducing explicit multipliers $\boldsymbol{\pi}_k \in \Re^{|\mathcal{N}|}$ for each commodity $k \in \mathcal{C}$, we obtain a problem, equivalent to VIP($\mathbf{t}$, $\widehat{F}_d$), which amounts to finding ($\mathbf{v}$, $\mathbf{w}$, $\mathbf{d}$, $\boldsymbol{\pi}$) such that

$$\mathbf{0}^{|\mathcal{L}|} \leqslant \mathbf{w}_k \perp \big(\mathbf{t}(\mathbf{v}) - \mathbf{E}^\top \boldsymbol{\pi}_k\big) \geqslant \mathbf{0}^{|\mathcal{L}|}, \quad k \in \mathcal{C}, \tag{52a}$$

$$\mathbf{E}\mathbf{w}_k = \mathbf{i}_k d_k, \quad k \in \mathcal{C}, \tag{52b}$$

$$\mathbf{g}(\mathbf{d}) = \boldsymbol{\pi}, \tag{52c}$$

$$\mathbf{v} - \sum_{k\in\mathcal{C}} \mathbf{w}_k = \mathbf{0}^{|\mathcal{L}|} \tag{52d}$$

holds. Fixing $\bar{\boldsymbol{\pi}}_k$ (and letting $\bar{\pi}_{pq} := \mathbf{i}_k^\top \bar{\boldsymbol{\pi}}_k$), and noting that (52b) is the dual optimality condition, the remaining problem is solved as follows. The demand $\mathbf{d}$ is given, through (52c), by the demand function at $\bar{\pi}$. The complementarity system (52a) is solved, in the separable case, by first defining, for each link $l = (i, j) \in \mathcal{L}$, $\bar{p}_{ij}(\boldsymbol{\pi}) := \text{maximum}_{k\in\mathcal{C}}\{\pi_{jk} - \pi_{ik}\}$ to be the largest node potential difference among the commodities. Let $\mathcal{C}_{ij}(\boldsymbol{\pi}) \subseteq \mathcal{C}$ denote the set of commodities for which $\pi_{jk} - \pi_{ik} = \bar{p}_{ij}(\boldsymbol{\pi})$ holds. Then, $v_{ij}(\boldsymbol{\pi}) := \text{maximum}\{0, t_{ij}^{-1}(\bar{p}_{ij}(\boldsymbol{\pi}))\}$ is the total link flow in link (i, j). The commodity link flows $w_{ij,k}$, $k \in \mathcal{C}$, can be taken as any distribution of the total flow $v_{ij}(\boldsymbol{\pi})$ onto the commodities in $\mathcal{C}_{ij}(\boldsymbol{\pi})$. It remains to define an updating rule for the vectors $\bar{\boldsymbol{\pi}}_k$ such that we approach also a solution to (52b). If this Lagrangian relaxation is utilized within a cyclic decomposition algorithm over OD pairs, where the subproblems are single-commodity network flow problems, then more efficient dual algorithms are available, since the duals then are differentiable.

4.3.2 Decomposition by linearization

A decomposition (relaxation) can also be achieved by manipulating the cost structure. Suppose, for example, that the link costs are fixed at their current values, and consider the resulting relaxation. Taking VIP($[\mathbf{t}, \mathbf{g}], \widehat{F}_d$) as the model for discussion, and the iterate $(\mathbf{v}^\tau, \mathbf{d}^\tau, (\boldsymbol{\pi}_k)_{k\in\mathcal{C}}^\tau)$ at iteration τ, fixing the link costs to $\mathbf{t}(\mathbf{v}^\tau)$ yields the following procedure for solving this relaxation of VIP($[\mathbf{t}, \mathbf{g}], \widehat{F}_d$): first, for each $k = (p, q) \in \mathcal{C}$, calculate a least-cost route, cf. (6a), whence the vectors $\bar{\boldsymbol{\pi}}_k^\tau$, $k \in \mathcal{C}$, $(\bar{\boldsymbol{\pi}}^\tau)_{(p,q)\in\mathcal{C}} = (\mathbf{i}_k^\top \bar{\boldsymbol{\pi}}_k^\tau)_{k\in\mathcal{C}}$ are given; second, calculate the demand through $\bar{\mathbf{d}}^\tau := \mathbf{g}(\bar{\boldsymbol{\pi}}^\tau)$; third, assign this demand to each OD pair's shortest route, giving rise to the link flows $\bar{\mathbf{w}}_k^\tau$, $k \in \mathcal{C}$, and $\bar{\mathbf{v}}^\tau = \sum_{k\in\mathcal{C}} \bar{\mathbf{w}}_k^\tau$. If link costs and demands are gradient mappings, we can perform a line search in the direction of $(\bar{\mathbf{v}}^\tau, \bar{\mathbf{d}}^\tau, (\bar{\boldsymbol{\pi}}_k)_{k\in\mathcal{C}}^\tau) - (\mathbf{v}^\tau, \mathbf{d}^\tau, (\boldsymbol{\pi}_k)_{k\in\mathcal{C}}^\tau)$, whence we have defined a partial linearization algorithm known as Evans' algorithm; whenever demand is inelastic, this is precisely the Frank–Wolfe algorithm.

The route information generated during the process of calculating shortest routes is badly utilized in line search algorithms, considering the effort in generating it. If, instead, we were to store the routes that carry flow or that are new and promising, and solve the equilibrium problem over those, we would in fact be addressing a restriction to the link–route formulation of the original model (a restricted master problem). As we have remarked earlier, this model form has an advantage in that the constraint structure can be effectively utilized; if we keep down the size of the route set, we also have a chance of avoiding the drawback of this formulation. Consider then the model VIP($\mathbf{t}, \widehat{F}$), where we have available a restriction $\widehat{\mathcal{R}} \subset \mathcal{R}$ of the route set. We recognize that the simplex structure can be put to good use, for instance by devising a decomposition over OD pairs or origins. This is achieved in a cyclic Newton-like algorithm, wherein we keep the flow from all but one group of users fixed, and

then utilize a diagonalized second-order approximation of the remaining objective function. At a given iteration where the flow is $\mathbf{h}^\tau$, we would, for one OD pair k, say, solve the following problem:

$$\underset{\mathbf{h}_k}{\text{minimize}} \sum_{r\in\widehat{\mathcal{R}}_k} \left\{ c_r(\mathbf{h}^\tau) h_r + \frac{b_r^\tau}{2\gamma_\tau} h_r^2 \right\}, \tag{53a}$$

$$\text{subject to} \quad \sum_{r\in\widehat{\mathcal{R}}_k} h_r \bar{d}_k, \tag{53b}$$

$$\mathbf{h}_k \geqslant \mathbf{0}^{|\widehat{\mathcal{R}}_k|}, \tag{53c}$$

where $0 < b_r^\tau \approx \partial c_r(\mathbf{h}^\tau)/\partial h_r$, and the value of $\gamma_\tau > 0$ is chosen such that the improvement made is the best possible; the resulting solution is taken to be $\mathbf{h}_k^{\tau+1}$. The above problem has a low complexity: solving for the optimal Lagrange multiplier for the equation can be done in $O(|\widehat{\mathcal{R}}_k|)$ time, that is, linear in the number of routes. Achieving low complexity is essential in the subproblems of an iterative algorithm, as the size of traffic problems can be very large.

A final remark on algorithms for the basic models is that it seems difficult to devise truly superlinearly convergent algorithms that utilize problem structure efficiently. One main reason is the presence of multiple commodities; for separable cost nonlinear single-commodity flow problems, superlinearly convergent algorithms are available, which also utilize the network property in an efficient way. However, the Gauss–Seidel algorithm – which is needed to coordinate the search in the multicommodity setting – is linearly convergent, at best. There is therefore still plenty of room for algorithmic improvements, in particular for new and clever ways of decomposing and coordinating the problem.

4.4 Merit functions and convergence for asymmetric models

When solving models involving cost and demand functions that are gradient mappings,[17] we have access to natural merit functions (cf. (15) or (16)), with which we can measure the progress of an algorithm and perform line searches; within this framework, the whole area of nonlinear optimization is open to exploration. When turning to nonintegrable models, these merit functions are no longer well defined, and progress toward an equilibrium must be measured by some artificial construct. One such function is the primal gap function given in (49) or (106), which can be calculated entirely in link flow space. As with most other merit functions used for variational inequality problems, its evaluation requires the solution of an approximation to the original problem. The computation of equilibria for nonseparable models are in fact much more complex than for separable ones:

[17] This is referred to as the 'integrable case'.

- *Decomposition*. In the integrable case, some of the most natural decompositions lead to low-complexity subproblems. In the nonintegrable case, however, they either become unavailable altogether, or they become more complex. For example, a decomposition over links will no longer be natural or even possible. Further, if the form of the cost and demand is complex, it also means that computing their values and derivatives will be more complex, and this will also affect the computational complexity of an algorithm.
- *Search direction → merit function*. In the integrable case, virtually every algorithm proposed is a feasible direction-based descent method which utilizes the natural objective in (15) or (16). Establishing convergence of such algorithms is relatively straightforward, as it is easy to satisfy the descent condition that the gradient of the merit function makes an obtuse angle with the search direction, and it does not rely on the monotonicity properties of the cost and demand functions. In the nonseparable case, however, the angle condition becomes much more difficult to establish for a suitable merit function. The merit function is intimately associated with the search direction, and in some cases it is even defined by the solution to the search direction-finding problem. That also means that, in contrast to the separable case, the merit function and its gradient depends on the search direction chosen. In order to derive convergence results, more stringent requirements on both the search direction and the step length must be enforced.

As an example of the complications that arise in the general, we mention that an algorithm that would mimic the Frank–Wolfe strategy by treating the cost function as if it were a gradient mapping, need not converge for asymmetric cost models.[18] However, it is straightforward to show that, whenever the 'Frank–Wolfe' direction is unique and the cost mapping is monotone, it constitutes a feasible descent direction for the primal gap function. If the set of Frank–Wolfe directions is not a singleton, it is yet possible to prove that there exists a member of that set that is gap-decreasing. Based upon this result, an implementable and convergent algorithm that operates in link flow space can be designed.

There exist several variational inequality algorithms that so far have not caught the attention of the traffic research community, and that converge under rather mild assumptions on the travel cost and demand mappings. For one, the projection algorithm, which takes an iterate $\mathbf{v}^\tau$ and maps it onto the next one through the formula

$$\mathbf{v}^{\tau+1} := \mathrm{Proj}_{\widehat{F}}\big(\mathbf{v}^\tau - \gamma_\tau \mathbf{t}(\mathbf{v}^\tau)\big), \quad \gamma_\tau > 0,$$

[18] In such an algorithm, line searches are performed based on first-order information, not the objective function.

where $\text{Proj}_{\widehat{F}}$ denotes the Euclidean projection onto $\widehat{F}$, can be supplied with special line searches that make it convergent for monotone cost mappings. (In its original statement it requires strong monotonicity or co-coercivity, and relies on the estimate of the cost mapping's strong monotonicity modulus and Lipschitz constant.) The class of extra-gradient algorithms involves two projections of the above type, and is convergent under even milder monotonicity requirements: pseudomonotonicity.

The most natural environment for these algorithms is the singly constrained restricted master problem in a route (column) generation algorithm, that is, a restriction of the feasible sets defined in (15) or (16) to a subset $\widehat{\mathcal{R}}$ of the routes. The reason why we favor this environment is that projections are easy to perform onto such sets, cf. (53).

An especially interesting avenue for constructing algorithms for nonstrictly monotone problems is to combine an algorithm of one's choice (and which perhaps requires strong monotonicity, such as the Jacobi algorithm) and the proximal point algorithm. It is defined thus: for the given cost function $\mathbf{t}$ in the variational inequality $\text{VIP}(\mathbf{t}, \widehat{F})$, one would add a multiple ($\gamma_\tau > 0$) times an affine cost of the form $\mathbf{v} \mapsto \mathbf{v} - \mathbf{v}^\tau$, if we work in link space. The resulting mapping $\mathbf{v} \mapsto \mathbf{t}(\mathbf{v}) + \gamma_\tau(\mathbf{v} - \mathbf{v}^\tau)$ is strongly monotone with a modulus of at least γ_τ, provided that the original cost function $\mathbf{t}$ is monotone. Therefore, if we have an algorithm which requires strong monotonicity, we apply it instead to this perturbed problem. At the (approximate) solution to the perturbed problem, we again perturb the original function at the new iterate, $\mathbf{v}^{\tau+1}$, and proceed, perhaps with a different value of γ.

Diagonalization is a favorite among heuristics for nonseparable problems. It amounts to, at an iteration, temporarily removing the dependency of a link's cost on the flow in other links. This is related to the Jacobi algorithm, and may suffer from quite poor convergence characteristics: if the correlation in reality is quite large, one can expect the algorithm to diverge, or at least converge slowly; on the other hand, if the cost dependence between links is mild, or local to a few links only, then convergence can be expected to be fast.

We close this section by mentioning that minimizing the dual gap function

$$\underset{\mathbf{x}\in X}{\text{maximum}}\ \mathbf{f}(\mathbf{y})^\top(\mathbf{x}-\mathbf{y})$$

can be achieved by solving the semiinfinite linear program to

$$\begin{aligned} &\underset{\mathbf{x}\in X, z\in\Re}{\text{minimize}}\ z \\ &\quad \text{subject to} \quad z \geqslant \mathbf{f}(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}), \quad \mathbf{y}\in X. \end{aligned} \tag{54}$$

This structure lends itself naturally to decomposition algorithms similar to the linear programming method of Dantzig and Wolfe, and to more recent interior point cutting plane methods.

4.5 Implementations and computational experience

Commercial implementations of equilibrium solvers are frequently based on the Frank–Wolfe method, or variants thereof. Although it has been known for quite some time that the Frank–Wolfe algorithm has the worst (sublinear) convergence rate among equilibrium algorithms, its high initial efficiency, small memory requirements, efficient handling of the underlying network structure and the natural interpretation behind its workings – it is reminiscent of heuristics that were devised already in the mid-1960s – have made it popular. Its counterpart for elastic demand models is Evans' algorithm, which has similar convergence characteristics: demand quickly converges, and once demand stabilizes, the algorithm's behavior is similar to Frank–Wolfe's.

Among the favored extensions of the Frank–Wolfe algorithm is the class of simplicial decomposition/column generation algorithms (see the discussion on the DSD algorithm above) which also are based on shortest route subproblems. While the simplicity of the subproblem is preserved, the results of the computations are better utilized, as the shortest routes are kept in memory and demand is optimally assigned onto them. This restricted master problem, which has the same form as that of the original problem except that the set $\mathcal{R}$ is replaced by the known subset $\widehat{\mathcal{R}}$, is then solved using a specialized, perhaps second-order, method. In later implementations of the DSD algorithm a diagonalized Newton algorithm is applied, which also efficiently detects routes with zero flow; these routes may be removed from the set $\widehat{\mathcal{R}}$ in order to save space. Other implementations use different versions of reduced gradient and gradient projection methods. As projection methods are available also for non-separable models, it is anticipated that the best versions can be transferred to solve also such more general models.

For a numerical illustration of two of the above discussions, let us consider the 'famous' Sioux Falls network, composed of 24 nodes, 76 links, and 528 OD pairs. For a MATLAB[19] implementation of the DSD and FW algorithms, Figure 10 illustrates how fast (in terms of CPU time) convergence occurs toward an equilibrium solution. The implementation of the DSD algorithm comes with several methods for solving the restricted master problems: two versions of the Goldstein–Levitin–Polyak gradient projection method (marked 'Gradient P.' in the figure), and a diagonalized Newton method. For the small-scale network representing Sioux Falls, the gradient projection method is faster, but for larger networks, Newton's method becomes relatively more favorable. Clearly, the Frank–Wolfe method is very slow for this problem, and its relative efficiency is even further reduced for larger problems.

A recent contribution to the field is the quasi-Newton algorithm of Bar–Gera. It utilizes the fact that the equilibrium flow from a given origin is acyclic

[19] The absolute CPU times should not be taken at face value; a corresponding C implementation can be two to three orders of magnitude faster.

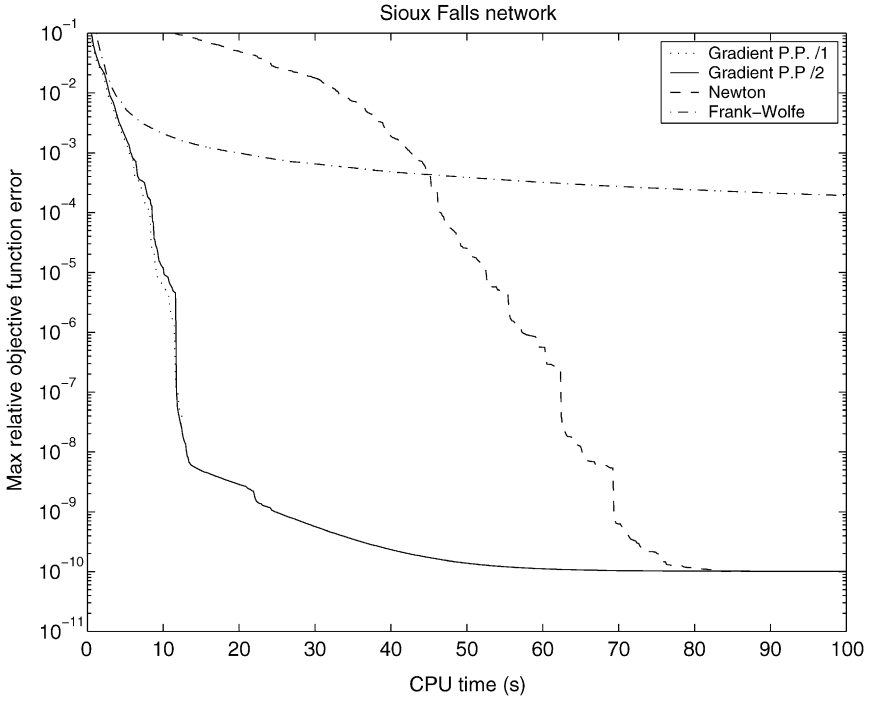


Fig. 10. The performance of DSD vs. FW on the Sioux Falls network.

(cf. Theorem 4(b)), and tries in a combinatorial fashion to determine the correct acyclic networks. For each origin and acyclic network, a quasi-Newton algorithm is implemented for finding the best single-commodity flow, based not only on link flows, but also on the flow through the nodes. The numerical experience so far is impressive, and it is believed that it will, if not replace, then at least become an alternative to the algorithms currently used in practice. The main drawback of this algorithm is that it has not yet been shown how to make it work for nonseparable models; see Section 5.4 for alternatives.

Somewhat surprisingly, the field of computational testing, in particular comparative ones, remains sparse. Comparisons are still quite often performed against the Frank–Wolfe algorithm; such results are not very enlightening, given the bad convergence behavior of the latter (see Figure 10).

Several interesting alternatives to the classical algorithms have never been tested, and the protocols of those that are tested on a regular basis would not pass the scrutiny of a numerical analysis specialist. The latter is not so surprising, considering that most of the algorithmic development relates to specific applications. The former is not so surprising either, given that most researchers in the field come from an area different from mathematical programming.

The low-complexity subproblems, like the shortest route problem and the quadratic knapsack problem (53), are also mathematical objects of interest.

Although there exist efficient algorithms for these problems, the fact that they have to be solved repeatedly leaves room for algorithms that are less efficient according to a worst-case criterion, but lend themselves better to reoptimization. As is the case for sorting, the worst-case behavior does not tell the whole story.

4.6 Bibliographical notes

The gap function defined by (49) is a measure of the distance to an equilibrium. Its negative is the directional derivative of the traffic equilibrium objective function ϕ (see (15)) along the direction toward a shortest route solution – which explains its role in the Frank–Wolfe algorithm. The gap function was originally considered by Zuhovickiĭ et al. (1969), and introduced into the field of transportation science by Murchland (1970); see also Auslender (1976).

Fixed-point type results concerning the algorithmic map (50) are based on contraction arguments (which are difficult to verify in practice), and can be found in Dunn (1973), Bruck (1975), and Baillon (1975). One reason we do not cover the fixed-point literature is that such formulations, through the use of problem mappings similar to (50), hide the salient structures of the underlying model. We prefer, departing somewhat from orthodoxy, to view such fixed-point problems not as traffic equilibrium models in their own right, but rather as algorithmic constructs.

The uncapacitated and link capacitated single-commodity flow problems discussed in Section 4.3, and algorithms for their solution, are reviewed in Kennington and Helgason (1980), Ahuja et al. (1993), Patriksson (1994b), Bertsekas (1998), Patriksson (2006).

The procedure following the statement (51) is taken from Larsson et al. (1997, 1999), and the procedure following the statement (52) stems from Patriksson (1994b, Section 4.3.2); both are covered in Patriksson (2006). The Evans and Frank–Wolfe methods were first implemented for traffic equilibrium models by Evans (1976), and Nguyen (1974) and LeBlanc et al. (1975), respectively.

The first successful attempt at building an efficient route-based column generation algorithm was made by Larsson and Patriksson (1992), who devised the *disaggregate simplicial decomposition* (DSD) framework. Although similar approaches had been proposed earlier, for example, in Leventhal et al. (1973) and Bertsekas and Gafni (1982), the restricted master problems were then deemed too difficult to solve, and route-based methods were therefore for many years considered inefficient. A few years later, Hearn et al. (1987) designed a similar algorithm where routes however are aggregated into all-or-nothing solutions, thereby restricting the number of variables in the restricted problems and at the same time reducing the potential speed of convergence for the respective route flows. The quadratic subproblem shown in (53) is the one preferred for the restricted problems, and is also used in the DSD algorithm. The convergence of this scaled gradient projection method has been

known for many years (e.g., Bertsekas, 1976). What makes it attractive is that the quadratic term can incorporate derivative information on route costs in order to enhance convergence, and also that the solution to (53) can be computed in linear time (Brucker, 1984). Although route-based formulations yield variational formulations that are not *strongly* monotone, convergence can yet be proved under the assumption that the *link* cost functions are strongly monotone (see Bertsekas and Gafni, 1982 and Zhu and Marcotte, 1996). (In the separable case, the problem's convexity is always enough to ensure convergence provided that a line search is performed in each iteration.)

For separable cost single-commodity problems, superlinearly convergent Newton-like methods that exploit network structure have been proposed by Best and Griffin (1975), Klincewicz (1983), Escudero (1986), Gafni and Bertsekas (1984); they are also overviewed in Patriksson (2006). In the multicommodity, nonseparable, case, a modified Newton method has been implemented by Marcotte and Guélat (1988) within the framework of simplicial decomposition. The linear convergence of the Gauss–Seidel algorithm is established in Bertsekas and Tsitsiklis (1989); the convergence rate depends of course on the extent to which the different variable components (that is, flow variables) interact in the objective function. Choosing the right decomposition – by OD pair, by origin, or even by grouping together some origins or OD pairs – may therefore be crucial for the practical convergence rate of the algorithm.

In the nonseparable case, the Frank–Wolfe strategy may lead to cycling (see, e.g., Hammond, 1984). It is interesting to note that column generation/simplicial decomposition algorithms are guaranteed to converge, as long as all routes are kept in memory, or if column dropping is done with care. Contributions to the theory and practice of such methods are found in Lawphongpanich and Hearn (1984) and Patriksson (1998). The existence of gap-decreasing Frank–Wolfe directions, and the design of a convergent framework built around such directions, is due to Marcotte (1986a).

The remainder of Section 4.4 discusses several ways in which convergence can be achieved by utilizing recent contributions to the field of variational inequality algorithms. General references for the development of line search methods for variational inequality problems until the late 1990s can be found in the monographs of Patriksson (1998) and Facchinei and Pang (2003b). Special line searches incorporated into the projection method are found in Bruck (1977) and Patriksson (1998). The class of extra-gradient algorithms has developed quite far since the original paper by Korpelevich (1977); recent contributions can be found in Iusem (1998), Konnov (2001), Wang et al. (2001), Solodov (2003). An implementation of such an algorithm in the context of network equilibrium can be found in Marcotte (1991). The proximal point algorithm was developed largely by Rockafellar (1976), following previous developments in the 1960s for more general problems. In that paper, an inexact solution of the regularized problems is already shown to be valid for monotone problems; more recent contributions, where the proximal point method is combined with projection methods, and where previous convergence rate analyses are im-

proved, are found in Solodov and Svaiter (1999a), Solodov and Svaiter (1999b), Solodov (2003). 'Diagonalization' is the favorite term in the transportation research literature (e.g., Sheffi, 1985) for the method known as 'Jacobi' to researchers in numerical analysis and mathematical programming. Cutting-plane methods for solving the Minty formulation or, equivalently, the semi-infinite linear program (54), have been proposed by Zuhovickiĭ et al. (1969) in a game-theoretical framework. It has been modified and adapted to the traffic assignment problem, using a column generation framework, by Nguyen and Dupuis (1984). The analytic center cutting plane method described in Goffin et al. (1997) may provide a viable alternative.

The contributions by Bar-Gera to the practice of solving separable, fixed demand problems are found in Bar-Gera (2002a, 2002b). Algorithms based on the DSD framework that have emerged since 1992 are found in, for example, Jayakrishnan et al. (1994) and Chen et al. (2002). The web site:

http://www.bgu.ac.il/∼bargera/tntp/

provides test problems and computational results. The reader may also consult

http://www-rocq.inria.fr/metalau/ciudadsim/.

Finally, the reoptimization issue has been addressed in Nguyen et al. (2002) and Pallottino and Scutellà (2003).

5 Solution algorithms: Variations

5.1 Multi-mode and multi-attribute models

In a multimodal context, decomposition is the strategy of choice for computing an equilibrium. If all modal flows are fixed, save one, the model reduces to a standard traffic equilibrium model which frequently involves a gradient mapping. At subiteration j of a major cycle τ, one solves the single-mode problem of finding the solution $\mathbf{w}^{j\tau}$ to the variational inequality

$$\mathbf{t}_m\big(\mathbf{v}^{1\tau}, \mathbf{v}^{j-1\tau}, \ldots, \mathbf{w}^{j\tau}, \mathbf{v}^{j+1\tau-1}, \ldots, v^{|\mathcal{M}|\tau-1}\big) \in N_{F_j}\big(\mathbf{w}^{j\tau}\big), \tag{55}$$

in the Gauss–Seidel variant and to the variational inequality

$$\mathbf{t}_m\big(\mathbf{v}^{1\tau-1}, \mathbf{v}^{j-1\tau-1}, \ldots, \mathbf{w}^{j\tau}, \mathbf{v}^{j+1\tau-1}, \ldots, v^{|\mathcal{M}|\tau-1}\big) \in N_{F_j}\big(\mathbf{w}^{j\tau}\big), \tag{56}$$

in the Jacobi (diagonalization) implementation. After the completion of an entire cycle, one updates the multi-attribute flow vector to

$$\mathbf{v}^{\tau+1} = \lambda \mathbf{v}^{\tau+1} + (1-\lambda)\mathbf{w}^{\tau}, \tag{57}$$

where $\lambda \in (0, 1)$ is a relaxation parameter. The convergence analysis of the above two schemes is available in classical numerical analysis textbooks and is reproduced in the transportation literature. It relies on the strong monotonicity and Lipschitz moduli of the cost mapping, two constants for which reliable

estimates may however be difficult to obtain; a trade-off has to be achieved between large values for λ that may yield to cycling, and small values that induce slow convergence.

The two-attribute structure is more interesting and lends itself to three lines of attacks. The first consists in discretizing the density function h and solving the resulting finite-dimensional problem by the decomposition approaches outlined above. A second strategy consists in solving a fixed point problem, either with respect to the total flow vector $\bar{\mathbf{v}}$ or the critical vector $\boldsymbol{\alpha}$, with the drawback that the choice of a suitable relaxation parameter may be difficult to estimate. We favor a third approach, viz. addressing the infinite-dimensional formulation (34) by a cost approximation method akin to the Frank–Wolfe linearization scheme. At iteration $\tau = 1$, one sets

$$\mathbf{v}^{\tau+1} = \mathbf{v}^{\tau} + \lambda_{\tau}\big(\mathbf{w}^{\tau} - \mathbf{v}^{\tau}\big), \tag{58}$$

where $\mathbf{w}^{\tau}$ is the solution to the infinite-dimensional parametric shortest paths problem. Since the information provided by the total flow vector $\bar{\mathbf{v}}$ is sufficient to recover the flow densities $\mathbf{v}(\alpha)$, the algorithm can also be defined in the finite-dimensional total flow space. If the conditions ensuring the validity of the optimization formulation (40) are fulfilled, λ can be determined by performing a line search along the direction $\mathbf{d}^{\tau} = \mathbf{w}^{\tau} - \mathbf{v}^{\tau}$. Otherwise, one may minimize, along $\mathbf{d}^{\tau}$, the primal gap function

$$\begin{aligned}
\operatorname{gap}(\mathbf{v}) &= \operatorname*{maximum}_{\mathbf{w}\in F} \big(\mathbf{t}(\bar{\mathbf{v}}) + \alpha\mathbf{f}(\bar{\mathbf{v}})\big)^{\top}(\mathbf{v} - \mathbf{w}) \\
&= \operatorname*{maximum}_{\mathbf{w}\in F} \int_0^{\bar{\alpha}} \big\langle \mathbf{t}(\bar{\mathbf{v}}) + \alpha\mathbf{f}(\bar{\mathbf{v}}), \mathbf{v}(\alpha) - \mathbf{w}(\alpha)\big\rangle \,\mathrm{d}\alpha \\
&= \operatorname*{maximum}_{\mathbf{w}\in F} \big\langle \mathbf{t}(\bar{\mathbf{v}}), \bar{\mathbf{v}} - \bar{\mathbf{w}}\big\rangle + \int_0^{\bar{\alpha}} \alpha\big\langle \mathbf{f}(\bar{\mathbf{v}}), \mathbf{v}(\alpha)\big\rangle \,\mathrm{d}\alpha \\
&\quad - \int_0^{\bar{\alpha}} \alpha\big\langle \mathbf{f}(\bar{\mathbf{v}}), \mathbf{w}(\alpha)\big\rangle \,\mathrm{d}\alpha.
\end{aligned}$$

The first term of this expression is the term found in the standard model, while the last two are the average costs associated with flow densities $\mathbf{v}$ and $\mathbf{w}$, respectively. If $\mathbf{f}$ is constant then the update of the latter quantity can be performed efficiently.

Without further assumptions on the problem's structure, the above described algorithm is a heuristic. However, the algorithm can be made globally convergent by imposing two simple conditions on the cost functions:

- *Invariance condition*: The mapping $\mathbf{f}$ is constant over its domain F.
- *Slope condition*: For any two distinct routes r_1 and r_2 there holds

$$\sum_{l\in r_1} f_l(\bar{\mathbf{v}}) \neq \sum_{l\in r_2} f_l(\bar{\mathbf{v}}), \quad \bar{\mathbf{v}} \in F. \tag{59}$$

The slope condition derives its name from the graph of the parametric LP, illustrated in Figure 5, whose solution yields the search direction $\mathbf{d}^\tau = \mathbf{w}^\tau - \mathbf{v}^\tau$. It states that the monetary cost associated with distinct routes are distinct. This condition can be enforced through a perturbation of the link costs, whenever the invariance condition holds.

Theorem 15. *Let the invariance and slope conditions hold. If* $\mathbf{t}$ *is monotone over* F*, then the link-flow solution* $\mathbf{v}^*$ *of the multi-attribute problem is unique, almost everywhere. Moreover, the path-flow solution is unique as well.*

The above results are surprising, as both conclusions may fail to hold for the standard model. Indeed, the first conclusion usually requires a strict monotonicity condition while the second fails if there exist more than one path flow vector compatible with a given link-flow solution. The situation is altogether different in the multi-attribute model, where the slope condition suffices to insure the uniqueness of the solution to a parametric LP. At equilibrium, every α-group is assigned to a unique path. While such a solution is extremal in the infinite-dimensional setting, the corresponding total flow vector will in all likelihood not coincide with an extreme point of the polyhedron F. Another consequence of the slope condition is that the search direction $\mathbf{d}^\tau$ induces a feasible but nonextremal total flow direction, relative to the polyhedron F. It is superior to an ordinary Frank–Wolfe direction in the sense that it is a descent direction for the gap function and induces convergence at a linear rate!

Theorem 16. *Let the invariance and slope conditions hold. Let* $\mathbf{t}$ *be strongly monotone* (*with modulus* b) *and Lipschitz continuous* (*with modulus* L) *over* F*. If the step size if fixed and satisfies* $0 < \lambda_\tau \equiv \lambda < \text{minimum}\{1, 2b/L\}$*, then the cost approximation algorithm converges geometrically to the equilibrium solution. More precisely*:

$$
\begin{aligned}
&g\big(\mathbf{v}^{\tau+1}\big) \leqslant (1-\lambda) g\big(\mathbf{v}^\tau\big),\\
&\big\|\bar{\mathbf{v}}^\tau - \bar{\mathbf{v}}^*\big\| = \mathcal{O}\big((1-\lambda)^{\tau/2}\big),\\
&\big\|\bar{\mathbf{w}}^\tau - \bar{\mathbf{w}}^*\big\| = \mathcal{O}\big((1-\lambda)^{\tau/2}\big)
\end{aligned}
$$

hold.

The last equality in the theorem states that the total flow vector $\bar{\mathbf{w}}^\tau$ corresponding to the solution $\mathbf{w}^\tau$ of the parametric shortest paths problem converges to the optimal total-flow equilibrium. Although the asymptotic rate of convergence is similar to that of the sequence $\bar{\mathbf{v}}^\tau$, the actual convergence is much slower, as observed empirically on large scale problems.

One drawback of the fixed step size scheme is that its validity rests on parameters b and L that might be difficult to estimate. This can be fixed by adopting a step size rule that obeys the Armijo condition with respect to the primal gap

function. Alternatively, if this solution is deemed to expensive computationally, the step size can be made to obey one of the following two rules:

$$\text{RULE I:} \quad 0 < \lambda_\tau < 1, \qquad \lim_{\tau \to \infty} \lambda_\tau = 0, \qquad \sum_{\tau=0}^{\infty} \lambda_\tau = +\infty;$$

$$\text{RULE II:} \quad \lambda_0 = 1, \qquad \lambda_{\tau+1} = \lambda_\tau - \frac{1}{2}\lambda_\tau^2.$$

In either case, global convergence is preserved, but asymptotic convergence may be sublinear.

Implementation. The main challenge in implementing the cost approximation algorithm lies in the numerical resolution of the parametric shortest paths problem. One may of course solve this parametric LP for its optimal solution by a parametric network simplex procedure, but this would prove costly on large networks involving thousands of paths, many of them irrelevant, and many others only active within a critical interval (α_{i-1}, α_i) of insignificant width. An alternative is to replace the density function h by a suitable discrete approximation. In practice, a 20-bar histogram yields results that are virtually indistinguishable from the exact solution. Might we then argue that the running time of a multi-attribute algorithm is roughly 20 times that of a standard model of similar size? No. We claim that the actual ratio is much smaller, as the convergence rate partially offsets the number of discrete problems (20) to be solved. Finally, one might ask what is the advantage of this approach over discretizing *h before* linearizing the functions **t** and **f**, a dilemma reminiscent of the Jacobi–Newton vs. Newton–Jacobi paradigm. Our answer is twofold. First, it is easier to monitor the convergence process through the control of a well-defined combinatorial subproblem. Second, let us not be shy about it: the analysis has some mathematical elegance.

5.2 Side constraints

The derivation of the side constrained traffic equilibrium model and its equilibrium characterization suggests the following algorithmic construct: guess a value of the multiplier vector $\boldsymbol{\beta}$, solve a standard traffic equilibrium model in terms of generalized link costs, and update the value of $\boldsymbol{\beta}$ based on a convergent algorithm for solving the Lagrangian dual problem. If the number of side constraints is relatively small, then this approach can be extremely efficient, since the subproblems are standard traffic equilibrium problems and the dual space has a small dimension.

Provided that the link travel cost functions are strictly increasing, the Lagrangian is strictly convex. Numerical experiments show however that, whenever an *augmented* Lagrangian algorithm is used for a link capacitated problem, convergence to a near-optimal, near-feasible solution is reached even more quickly. Starting from a near-optimal and near-feasible solution, it is then

relatively straightforward to generate a *primal feasible* and near-optimal solution by means of specialized graph search techniques working in a residual graph.

5.3 Strategic model

In this section, we develop a strategy for solving the basic model. These ideas can be extended to the dynamic and priority models. The challenge in finding an equilibrium solution is threefold:

(i) The cost mapping $\mathbf{c}$ is not available in closed form.
(ii) $\mathbf{c}$ is neither differentiable nor monotone.
(iii) The entire information pertaining to the strategic flow $\mathbf{x}$ must be preserved at every iteration.

The second issue implies that only heuristic methods based on first-order information are implementable. One such proposal is based on the solution of a linear program; at iteration $\tau + 1$, one sets

$$\mathbf{x}^{\tau+1} = \mathbf{x}^{\tau} + \lambda_{\tau}\big(\mathbf{y}^{\tau} - \mathbf{x}^{\tau}\big),$$

where $\lambda_{\tau} \in [0, 1]$ and

$$\mathbf{y}^{\tau} \in \arg\operatorname*{minimum}_{\mathbf{y}\in X} \mathbf{c}\big(\mathbf{x}^{\tau}\big)^{\top}\mathbf{y},$$

the vector of best strategic responses to $\mathbf{x}^{\tau}$, is obtained by solving Bellman's dynamic programming equations, for every destination node q:

$$\omega_j^{s^*} = \begin{cases} \infty, & \text{if } j > q, \\ 0, & \text{if } j = q, \\ \sum_{k\in E_j^{s^*}} \pi_{jk}^{s^*}\big(t_{jk} + \omega_k^{s^*}\big), & \text{if } j < q, \end{cases} \tag{60}$$

and assigning all strategic flow $\mathbf{y}^{\tau}$ to strategy s^*. In the above expression, $E_j^{s^*}$ denotes the preference order of strategy s^*, $\pi_{jk}^{s^*}$ is the probability of accessing node k from node j, using s^*, t_{jk} is the traversal time of link (j, k) and $\omega_k^{s^*}$ is the cost-to-go function. Since the vector $\boldsymbol{\pi}$ is a by-product of the loading process, [20] the recursion will be well-defined once the preference orders are set. At a given node, one may of course consider all possible such orders. However, under the FIFO queue discipline, one can prove the optimality of the following greedy approach:

$$E_j^{s^*} \in \arg\operatorname*{minimum}_{E_j^s} \sum_{k\in E_j^s} \pi_{jk}^{s}\big(t_{jk} + \omega_k^{s^*}\big). \tag{61}$$

[20] This result also holds for strategies that have not been considered yet, i.e., strategies that carry no flow.

The time required to compute $\mathbf{c}(\mathbf{x})$ favors the use of a preselected sequence of step sizes, such as the harmonic sequence $\lambda_\tau = 1/\tau$. An alternative which proves efficient in practice is to weight the harmonic step size $1/\tau$ by the nonnegative vector $\mathbf{c}(\mathbf{x}^\tau) - \mathbf{c}(\mathbf{y}^\tau)$, which is a measure of departure from equilibrium or, in layperson terms, a measure of user dissatisfaction. The third point is best addressed by resorting to restriction (simplicial decomposition) techniques.

5.4 *Nonadditive route costs*

Algorithms for nonadditive traffic equilibrium models need explicit route information; a simplicial decomposition/route (column) generation algorithm à la DSD therefore seems natural. In the case of the above cost model, the only place where some complications arise is in the generation of new, profitable routes. The Wardrop conditions suggest the following extension of the shortest route problem, for an OD pair $k \in \mathcal{C}$, at some fixed link volume $\bar{\mathbf{v}}$:

$$\underset{r \in \mathcal{R}_k}{\text{minimize}} \sum_{l \in \mathcal{L}} \lambda_{lr} t_l(\bar{\mathbf{v}}) + \alpha_r(f_r).$$

If the current solution is not in equilibrium, it can be improved through the reallocation of flow onto shortest routes, exactly as in Frank–Wolfe's linearization method. Unfortunately, the above shortest route problem cannot be solved using Dijkstra's algorithm, since each link possesses two attributes: travel time and money outlay; in particular, the dynamic programming principle – that subroutes are optimal if the entire routes are – does not apply. Instead, more complex multilabel shortest route methods must be used.

Suppose that $f_r = \sum_{l \in \mathcal{L}} \lambda_{lr} f_{lr}$, so that the monetary outlay on the route is the sum of link tolls. Then we can devise an algorithm in which node labels incorporate both accumulated travel times and monetary outlays, from a specific node on to the destination nodes. Domination tests (based on the principle of Pareto optimality) are used to keep down the number of labels stored, and those remaining at the terminal node are compared with respect to the function ν_k to determine the best route. The corresponding extension of the DSD algorithm works well for this problem.

5.5 *Bibliographical notes*

Algorithms for the bi-attribute model have been analyzed in Marcotte and Zhu (1997). An implementation based on the parametric network simplex method is described in Marcotte et al. (1996) in the congestion-free case.

The side constrained model discussed in Section 5.2 had previously been considered algorithmically mainly for the link capacitated case, typically through penalty methods. The algorithm in Larsson and Patriksson (1995) (whose literature section traces the history of such methods within the context of transportation science) combines an augmented Lagrangian method with

the DSD algorithm used for the subproblem. One of the few algorithms for the general side constrained model is found in Larsson et al. (2004), which combines Lagrangian relaxation and column generation.

The implementation of the strategic model is discussed in Marcotte et al. (2004), while numerical results on the nonadditive model are found in Larsson et al. (2002).

6 Optimization in a user equilibrium context

Equilibrium models are by construction descriptive, passive mathematical objects that can be used to assess the impact of managerial policies. Ideally, a design model should be able to integrate equilibrium equations within the overall control process. In this section, we address the design problem both from the local and global perspectives. In this respect, the first subsection provides sensitivity results for the traffic equilibrium problem, while the next two subsections are concerned with the problem of imposing tolls either for controlling flow patterns through the network, or for raising revenues.

6.1 Stability and sensitivity of traffic equilibria

Although Braess' paradox, which was discovered in the late 1960s, prompted a study into the qualitative behavior of a traffic equilibrium under varying conditions, a thorough quantitative study, in the form of a variational and differential analysis, is much more recent. This section traces the main variational characteristics of a traffic equilibrium, and shows when the equilibrium is (directionally) differentiable as a function of input data; the results cited improve quite substantially on previous analyses that relied heavily on the Implicit Function Theorem and are therefore only applicable in very restrictive settings.

6.1.1 Introduction

The basis of our sensitivity analysis is a result which is stated for a general variational inequality problem over a polyhedron and with a differentiable mapping, $\mathbf{f}: \Re^d \times \Re^n \to \Re^n$ in the parameters $\boldsymbol{\rho} \in \Re^d$ and variables $\mathbf{x} \in \Re^n$: find $\mathbf{x}^* \in X$ such that

$$\mathbf{f}(\boldsymbol{\rho}, \mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geqslant 0, \quad \mathbf{x} \in X, \tag{62}$$

where $X \subseteq \Re^n$ is a polyhedral set.[21] We let $\mathbf{S}: \Re^d \rightrightarrows 2\Re^d$ denote the mapping that assigns to each vector $\boldsymbol{\rho} \in \Re^d$ the set $\mathbf{S}(\boldsymbol{\rho})$ of solutions to this problem.

[21] That the set X is polyhedral is crucial. It is however possible to take into account nonlinear constraints by reformulating the problem such that Lagrange multipliers are explicitly used.

Letting $\boldsymbol{\rho} = \boldsymbol{\rho}^*$ be the current value of the parameter vector, we are interested in the direction of change of the solution $\mathbf{x}^*$ as $\boldsymbol{\rho}^*$ is perturbed along a direction $\boldsymbol{\rho}'$. This directional derivative of $\mathbf{S}$ is the solution to an auxiliary variational inequality, which has the following form: find $\mathbf{x}' \in K$ such that

$$\mathbf{r}(\boldsymbol{\rho}', \mathbf{x}')^\top (\mathbf{x} - \mathbf{x}') \geqslant 0, \quad \mathbf{x} \in K, \tag{63a}$$

where

$$K := T_X(\mathbf{x}^*) \cap \mathbf{f}(\boldsymbol{\rho}^*, \mathbf{x}^*)^\perp \tag{63b}$$

and

$$\mathbf{r}(\boldsymbol{\rho}', \mathbf{x}') := \nabla_{\boldsymbol{\rho}} \mathbf{f}(\boldsymbol{\rho}^*, \mathbf{x}^*)\boldsymbol{\rho}' + \nabla_{\mathbf{x}} \mathbf{f}(\boldsymbol{\rho}^*, \mathbf{x}^*)\mathbf{x}'. \tag{63c}$$

We let $\mathbf{DS}(\boldsymbol{\rho}^*|\mathbf{x}^*) : \Re^d \rightrightarrows \Re^d$ denote the mapping that assigns to each perturbation $\boldsymbol{\rho}' \in \Re^d$ the set $\mathbf{DS}(\boldsymbol{\rho}^*|\mathbf{x}^*)(\boldsymbol{\rho}')$ of solutions to this problem. The set K above is referred to as the *critical cone*, and denotes the set of variations around $\mathbf{x}^*$ that, roughly speaking, retains feasibility and optimality to the first order. T_X denotes the tangent cone to X, which means that if X is defined by linear constraints, we have that

$$\begin{aligned} X &= \left\{\mathbf{x} \in \Re^n \mid \mathbf{A}\mathbf{x} \geqslant \mathbf{b};\, \mathbf{B}\mathbf{x} = \mathbf{d}\right\} \\ &\Longrightarrow \quad T_X(\mathbf{x}^*) = \left\{\mathbf{z} \in \Re^n \mid \overline{\mathbf{A}}\mathbf{z} \geqslant \mathbf{0};\, \mathbf{B}\mathbf{z} = \mathbf{0}\right\}, \end{aligned}$$

where $\overline{\mathbf{A}}$ consists of the rows $\mathbf{A}_i$ of $\mathbf{A}$ corresponding to the binding inequality constraints at $\mathbf{x}^*$, that is, the indices i with $\mathbf{A}_i\mathbf{x}^* = b_i$. Further, for any vector $\mathbf{z} \in \Re^n$, $\mathbf{z}^\perp := \{\mathbf{y} \in \Re^n \mid \mathbf{z}^\top \mathbf{y} = 0\}$ is the orthogonal subspace associated with the vector $\mathbf{z}$. The mapping $\mathbf{r}$ is a linearization of $\mathbf{f}$ around $(\boldsymbol{\rho}^*, \mathbf{x}^*)$; it is an affine mapping in $\mathbf{x}'$.

Suppose now that $\mathbf{f}(\boldsymbol{\rho}, \cdot)$ is monotone on X around $\boldsymbol{\rho} = \boldsymbol{\rho}^*$, and that the parameterization is such that the rank of the matrix $\nabla_{\boldsymbol{\rho}} \mathbf{f}(\boldsymbol{\rho}^*, \mathbf{x}^*)$ is equal to n. (The latter result can always be fulfilled by including dummy parameters.) We say that the mapping $\mathbf{S}$ is *strongly regular* at $\boldsymbol{\rho}^*$ if $\mathbf{S}$ is single valued and Lipschitz continuous on some neighborhood of $\boldsymbol{\rho}^*$. Then,

$$\mathbf{S} \text{ is strongly regular at } \boldsymbol{\rho}^* \tag{64a}$$

$$\Longleftrightarrow \quad \mathbf{DS}(\boldsymbol{\rho}^*|\mathbf{x}^*) \text{ is single valued.} \tag{64b}$$

Moreover, the unique solution $\mathbf{x}' \in \Re^n$ to (63) is the *directional derivative* of the solution $\mathbf{x}^*$ to (62) at $\boldsymbol{\rho}^*$, in the direction of $\boldsymbol{\rho}'$. A sufficient condition for the property (64b) to hold is that

$$\nabla_{\mathbf{x}} \mathbf{f}(\boldsymbol{\rho}^*, \mathbf{x}^*) \text{ is positive definite on } (K - K). \tag{65}$$

We refer to this as a sufficient *second-order* condition. A result stronger than directional differentiability can also be obtained under additional assumptions.

Indeed, under strong regularity,

$$\mathbf{S} \text{ is differentiable at } \boldsymbol{\rho}^* \tag{66a}$$

$$\iff \quad \mathbf{DS}(\boldsymbol{\rho}^*|\mathbf{x}^*)(\boldsymbol{\rho}') \in -K, \quad \boldsymbol{\rho}' \in \Re^d. \tag{66b}$$

Moreover, if the critical cone K is a subspace, that is, if $K = K \cap (-K)$, then the gradient can be represented as

$$\nabla_{\boldsymbol{\rho}} \mathbf{x}(\boldsymbol{\rho}^*) = -\mathbf{Z}\big[\mathbf{Z}^\top \nabla_{\mathbf{x}} \mathbf{f}(\boldsymbol{\rho}^*, \mathbf{x}^*)\mathbf{Z}\big]^{-1} \mathbf{Z}^\top \nabla_{\boldsymbol{\rho}} \mathbf{f}(\boldsymbol{\rho}^*, \mathbf{x}^*), \tag{67}$$

for any $n \times \ell$ matrix $\mathbf{Z}$ such that $\mathbf{Z}^\top\mathbf{Z}$ is nonsingular and $\mathbf{z} \in K \cap (-K)$ if and only if $\mathbf{z} = \mathbf{Z}\mathbf{y}$ for some $\mathbf{y} \in \Re^\ell$, where ℓ is the dimension of $K \cap (-K)$. This differentiability result is a kind of implicit function theorem; the relationships in (64) and (66) show how the implicit function theorem naturally extends to more general cases.

The latter property has been invoked in many sensitivity analysis frameworks. That is, they hinge on being able to invoke the implicit function theorem. Unfortunately, not only does the property $\mathbf{DS}(\boldsymbol{\rho}^*|\mathbf{x}^*)(\boldsymbol{\rho}') \in -K$ fail to hold in many cases (cf. below), but also there may not exist a nonsingular matrix of the kind referred to above.

6.1.2 Sensitivity analysis of separable traffic equilibrium

For simplicity of presentation, we focus on the case of separable cost and demand functions. First, we cast the VIP (8) in the sensitivity framework. Let

$$\mathbf{x} = \begin{pmatrix} \mathbf{h} \\ \mathbf{v} \\ \mathbf{d} \end{pmatrix}; \qquad \mathbf{f}(\boldsymbol{\rho}, \mathbf{x}) = \begin{pmatrix} \mathbf{0}^{|\mathcal{R}|} \\ \mathbf{t}(\boldsymbol{\rho}, \mathbf{v}) \\ -\boldsymbol{\xi}(\boldsymbol{\rho}, \mathbf{d}) \end{pmatrix}; \qquad X = \Re_+^{|\mathcal{R}|} \times \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|}.$$

Then, we can identify the sensitivity problem through the following identifications:

$$K = \left\{ \begin{pmatrix} \mathbf{h}' \\ \mathbf{v}' \\ \mathbf{d}' \end{pmatrix} \in \Re^{|\mathcal{R}|} \times \Re^{|\mathcal{L}|} \times \Re^{|\mathcal{C}|} \,\middle|\, \boldsymbol{\Gamma}^\top \mathbf{h}' = \mathbf{d}';\ \mathbf{v}' = \boldsymbol{\Lambda}\mathbf{h}';\ \mathbf{h}' \in H' \right\},$$

where

$$H' = \left\{ \mathbf{h}' \in \Re^{|\mathcal{R}|} \,\middle|\, \begin{array}{ll} h'_r \text{ free}, & \text{if } h^*_r > 0, \\ h'_r \geqslant 0, & \text{if } h^*_r = 0 \text{ and } c_r(\boldsymbol{\rho}^*, \mathbf{h}^*) = \pi^*_k, \\ h'_r = 0, & \text{if } h^*_r = 0 \text{ and } c_r(\boldsymbol{\rho}^*, \mathbf{h}^*) > \pi^*_k, \\ [r \in \mathcal{R}_k, k \in \mathcal{C}] & \end{array} \right\},$$

and

$$\mathbf{r}(\boldsymbol{\rho}', \mathbf{x}') = \begin{pmatrix} \mathbf{0}^{|\mathcal{R}|} \\ \nabla_{\boldsymbol{\rho}} \mathbf{t}(\boldsymbol{\rho}^*, \mathbf{v}^*)\boldsymbol{\rho}' + \nabla_{\mathbf{v}} \mathbf{t}(\boldsymbol{\rho}^*, \mathbf{v}^*)\mathbf{v}' \\ -[\nabla_{\boldsymbol{\rho}} \boldsymbol{\xi}(\boldsymbol{\rho}^*, \mathbf{d}^*)\boldsymbol{\rho}' + \nabla_{\mathbf{d}} \boldsymbol{\xi}(\boldsymbol{\rho}^*, \mathbf{d}^*)\mathbf{d}'] \end{pmatrix}.$$

By the monotonicity and separability of $\mathbf{t}$ and $-\boldsymbol{\xi}$, the resulting 'sensitivity variational inequality' (63) can be equivalently written as the convex quadratic

optimization problem to

$$\begin{aligned}\underset{(\mathbf{v}',\mathbf{d}',\mathbf{h}')}{\text{minimize}}\ \phi'(\mathbf{v}',\mathbf{d}') := &\left[\nabla_{\boldsymbol{\rho}}\mathbf{t}(\boldsymbol{\rho}^*,\mathbf{v}^*)\boldsymbol{\rho}'\right]^\top\mathbf{v}' \\ &+\frac{1}{2}\sum_{l\in\mathcal{L}}\frac{\partial t_l(\boldsymbol{\rho}^*,v_l^*)}{\partial v_l}(v_l')^2 \\ &-\left[\nabla_{\boldsymbol{\rho}}\boldsymbol{\xi}(\boldsymbol{\rho}^*,\mathbf{d}^*)\boldsymbol{\rho}'\right]^\top\mathbf{d}' \\ &-\frac{1}{2}\sum_{k\in\mathcal{C}}\frac{\partial \xi_k(\boldsymbol{\rho}^*,d_k^*)}{\partial d_k}(d_k')^2, \qquad (68a)\end{aligned}$$

$$\text{subject to}\quad \boldsymbol{\Gamma}^\top\mathbf{h}' = \mathbf{d}', \qquad (68b)$$

$$\mathbf{v}' = \boldsymbol{\Lambda}\mathbf{h}', \qquad (68c)$$

$$\mathbf{h}' \in H'. \qquad (68d)$$

The sensitivity problem is closely related to the original model, with two notable differences: the link cost and demand functions are replaced by their linearizations, and the sign restrictions on $\mathbf{h}$ are replaced by individual restrictions on the route flow perturbations h_r' that depend on whether the route in question was used at equilibrium or not, cf. the set H'. Although the appearance of H' depends on the choice of route flow solution $\mathbf{h}^*$, it is an interesting fact that the possible choices of $\mathbf{v}'$ in K do *not*; this is a general consequence of aggregation, and amounts to the possibility of (essentially) eliminating route flow variables from explicit consideration; this is only possible because of the special connection between route and link flow variables. Finally, we see that the resemblance to the original problem implies that the sensitivity variational inequality problem can be solved using software similar to those for the original traffic equilibrium model, provided of course that route flow information can be extracted.

Assumption 17 (Properties of the network model).

(a) For each $l \in \mathcal{L}$, the link travel cost function $t_l(\cdot,\cdot)$ is continuously differentiable, and strictly increasing in its second argument.
(b) For each $k \in \mathcal{C}$, the demand function $g_k(\cdot,\cdot)$ is continuously differentiable, nonnegative, upper bounded, and strictly decreasing in its second argument. The function $g_k(\rho,\cdot)$ is therefore invertible, and has a single-valued inverse, $\xi_k(\rho,\cdot)$, which also is continuously differentiable and strictly decreasing.

Theorem 18 (Sensitivity of separable traffic equilibrium problems). *Let Assumption 17 hold, and consider an arbitrary vector* $\boldsymbol{\rho}^* \in \Re^d$. *Then, the solution* $(\mathbf{v}^*,\mathbf{d}^*)$ *to* (16) *is unique, and so are the* (*negative*) *travel cost entities* $(\mathbf{s}^*,\boldsymbol{\pi}_-^*) = -(\mathbf{t}(\boldsymbol{\rho}^*,\mathbf{v}^*),\boldsymbol{\xi}(\boldsymbol{\rho}^*,\mathbf{d}^*))$.

Assume that the link travel cost function $\mathbf{t}(\boldsymbol{\rho}^*, \cdot)$ *is such that*

$$\frac{\partial t_l(\boldsymbol{\rho}^*, v_l^*)}{\partial v_l} > 0, \quad l \in \mathcal{L}. \tag{69}$$

Assume further that the demand function $\mathbf{g}(\boldsymbol{\rho}^*, \cdot)$ *is such that*[22]

$$\frac{\partial g_k(\boldsymbol{\rho}^*, \pi_k^*)}{\partial \pi_k} < 0, \quad k \in \mathcal{C}. \tag{70}$$

Then, in the solution to (68)*, the values of the link flow and demand perturbation* $\mathbf{v}'$ *and* $\mathbf{d}'$ *are unique; therefore, the value* $\mathbf{v}'$ (*respectively,* $\mathbf{d}'$) *is the directional derivative of the equilibrium link flow* (*respectively, demand*)*, at* $\boldsymbol{\rho}^*$*, in the direction* $\boldsymbol{\rho}'$.

6.1.3 Limitations of the classical sensitivity framework

Sensitivity analyses that extend standard techniques for nonlinear optimization and variational inequality problems are widespread, and have been adopted by several researchers over the past fifteen years, particularly as a subroutine in the solution of more complex problems of a bilevel nature (see, for example, Section 6.4). These analyses have however a drawback: they often are inapplicable because of the strong assumptions underlying their validity. Part of the problem lies in the utilization of the implicit function theorem, which requires the problem to be expressed as a system of equations; this leads to a strict complementarity type of condition, which may fail to be in force at differentiable points. Additional requirements pose, for example, conditions on the topology of the network itself, and are not necessary either. It appears that some researchers are not aware of these limitations, and of the existence of more widely applicable analyses. This justifies devoting some space to examples where this 'classical' sensitivity framework fails, while the techniques of the previous section still apply.

Example 19 (A nonstrictly complementary example). The classical analysis requires that the solution be strictly complementary. In our context this means that there must exist some equilibrium route flow solution $\mathbf{h}^*$ which has the property that for every route in $\mathcal{R}$, $h_r^* > 0$ holds if and only if route r is a shortest route given the equilibrium travel costs. The definition in the classical work is however based on the total link flows and an aggregated 'node price', common to all OD pairs, which may not exist (see, for example, the numerical example in Section 2).

To show that strict complementarity is not necessary for differentiability (as we have already seen from the above analysis), we consider the network depicted in Figure 11. It involves two OD pairs, $(1, 2)$ and $(4, 2)$, with a fixed

[22] The two derivative conditions (69) and (70) imply that the functions $t_l(\boldsymbol{\rho}^*, \cdot)$ and $-g_k(\boldsymbol{\rho}^*, \cdot)$ are strictly increasing.

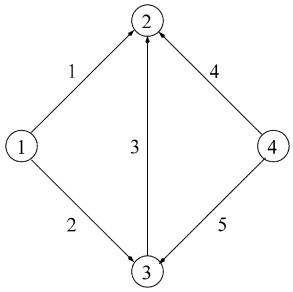


Fig. 11. Network for the first example.

and unperturbed demand of 2 and 1 units of flow, respectively. The link cost functions are given by

$$t_1(v_1, \rho) := 2v_1 + \rho; \qquad t_2(v_2) := v_2; \qquad t_3(v_3) := 1;$$

and

$$t_4(v_4) := v_4 + 2; \qquad t_5(v_5) = v_5.$$

We have four routes: {1}, {2, 3}, {4}, and {5, 3}, two for each OD pair.

If $\rho^* = 0$, the unperturbed traffic equilibrium solution is $\mathbf{v}^* = (1, 1, 1, 1, 1)^\top$ and the route flow $\mathbf{h}^* = (1, 1, 0, 1)^\top$ is unique. We observe that the travel cost on route 3 is equal to 2, as is the case for route 4, so this equilibrium solution is not strictly complementary.

In order to check whether the solution $\mathbf{v}^*$ is nevertheless differentiable at $\rho^* = 0$, we solve the sensitivity problem for both $\rho' := 1$ and $\rho' := -1$. For $\rho' = 1$, we obtain the following unique solution to the sensitivity problem, thus, being the directional derivative of $\mathbf{v}^*$ with respect to the direction $\rho' = 1$ at $\rho^* = 0$: $\mathbf{v}' = (-\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0)^\top$. The effect of perturbing link 1's cost, such that it becomes more expensive, is that of sending flow along the cycle $\{-1, 2, 3\}$, where the minus sign reflects that flow is sent backward on link 1. When solving the sensitivity problem for $\rho' := -1$, we obtain the directional derivative $\mathbf{v}' = (\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, 0, 0)^\top$, that is, the negative of the directional derivative of $\mathbf{v}^*$ in the direction of $\rho' := 1$. This proves that the directional derivative mapping is linear, and thus that the derivative of $\mathbf{v}^*$ with respect to ρ' at $\rho^* = 0$ equals $\mathrm{d}\mathbf{v}^*/\mathrm{d}\rho = (-\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0)^\top$.

Example 20 (Second example). A key assumption of the classical analysis is that the travel cost function $\mathbf{t}(\boldsymbol{\rho}, \cdot)$ be strongly monotone in a neighborhood of the equilibrium solution. This is clearly not needed, as we have shown above.

More serious, however, is the presence of a condition on the topology of the graph representing the traffic network. Suppose that we were to limit the discussion to a subgraph of $\mathcal{G}$ in which only the links carrying positive flow are included. In this subnetwork, the analysis must be performed from an equilibrium route flow solution that is an extreme point of H (the routes using the

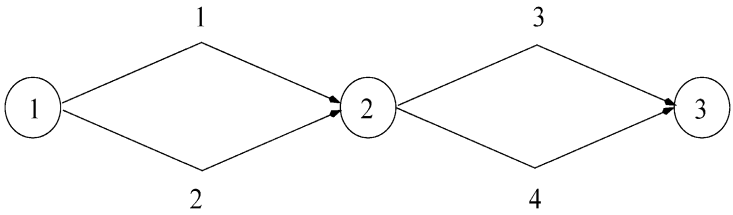


Fig. 12. Network for the second example.

above-stricken zero-flow links also having been stricken) which has exactly as many routes with positive flow as the rank of the matrix $[\boldsymbol{\Lambda}_+^\top | \boldsymbol{\Gamma}_+]$ (the + sign indicates that we have eliminated the zero-flow routes, as discussed). The rank of this matrix is never larger than the number of links with positive flow at $\mathbf{v}^*$ plus $|\mathcal{C}|$. Although the choice of the route flow solution is immaterial, it must be extremal, as explained above. The resulting formula then takes the form

$$\begin{pmatrix} \nabla_{\boldsymbol{\rho}} \mathbf{h}_+ \\ \nabla_{\boldsymbol{\rho}} \boldsymbol{\pi} \end{pmatrix} = \begin{pmatrix} \nabla_{\mathbf{h}} \mathbf{c}_+(\boldsymbol{\rho}^*, \mathbf{h}_+^*) & -\boldsymbol{\Lambda}_+^\top \\ \boldsymbol{\Lambda}_+ & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} -\nabla_{\boldsymbol{\rho}} \mathbf{c}_+(\boldsymbol{\rho}^*, \mathbf{h}_+^*) \\ \nabla_{\boldsymbol{\rho}} \mathbf{g}(\boldsymbol{\rho}^*) \end{pmatrix}. \tag{71}$$

Unfortunately, it is not difficult to construct examples where the solution is differentiable but where the matrix condition fails, so that the formula (71) breaks down. Consider the network shown in Figure 12, involving a single OD pair, (1, 3), and a fixed demand of 2 flow units. Set the link cost functions to

$$t_1(v_1, \rho) = v_1 + \rho; \qquad t_2(v_2) = v_2; \qquad t_3(v_3) = v_3; \qquad t_4(v_4) = v_4.$$

In this example, we have four routes: {1, 3}, {1, 4}, {2, 3}, and {2, 4}.

If $\rho^* = 0$, the unperturbed traffic equilibrium solution is $\mathbf{v}^* = (1, 1, 1, 1)^\top$. We can easily see that the solution is differentiable, and even strictly complementary. The derivative with respect to ρ at ρ^* is $(-\frac{1}{2}, 0, \frac{1}{2}, 0)^\top$. This result is intuitive: if the value of ρ increases, then the flow on link 1 should decrease, whence the flow on link 2 must increase by the same amount. If, on the other hand, the value of ρ decreases, the reverse should happen.

Consider then the workings of the classical formula (71). We obviously fulfill the strong monotonicity conditions on the travel cost function. Since all links carry flow at equilibrium, we need not remove any links or routes when considering the sensitivity analysis problem. We last try to comply with the linear independence condition, by choosing the right equilibrium route flow solution. Note then that

$$\left[\boldsymbol{\Lambda}^\top | \boldsymbol{\Gamma}\right] = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix},$$

which has rank 3. So, we should find a route flow solution, $\mathbf{h}^*$, in which exactly 3 routes have a positive flow. This is however impossible; the only alternatives

are 2 or 4. To see why, let us suppose that the flow on the first route, $\{1, 3\}$, is $\alpha \in [0, 1]$. Then, the flows on routes $\{1, 4\}$ and $\{2, 3\}$ must both be $1 - \alpha$, in order to comply with the total flow on the links. This implies that the flow on route $\{2, 4\}$ is α and that, for any value of $\alpha \in [0, 1]$, the number of routes carrying nonzero flow is either 2 or 4. We can therefore not comply with the matrix condition stated, and the classical formula (71) fails, even though the gradient does exist.

Note that the technique of Section 6.1.2 carries over to the asymmetric case. This approach yields the recommended source for sensitivity information, not only because it is widely applicable, but also because the sensitivity analysis problem, being closely related to that of the original equilibrium problem, can be solved using similar computational tools. This is especially important when relying on sensitivity information for an extended traffic model, for example of the bilevel type, where many such analyses must be performed (cf. Section 6.4). The classical formula may then fail, either because of the topology being 'wrong' or for lack of strict complementarity (and the latter *will* typically be the case at a stationary point of a bilevel program); in either case, the formula breaks down, regardless of the existence, or not, of a gradient. It is common to claim that the possibility of the nondifferentiability of a given point in the space of $\boldsymbol{\rho}$ can be ignored because of Rademacher's theorem, which essentially states that a locally Lipschitz continuous function is differentiable everywhere except possibly on a set with a zero (Lebesgue) measure. This type of argument ignores that several of the most interesting points to look at *are* points of the latter category; for example, optimal solutions to bilevel programs and, more generally, MPECs (Mathematical Program with Equilibrium Constraints) problems, are indeed extremal.[23]

6.2 *Inducing system-optimal flows through link tolls*

Tolls, either physical or virtual, may be used not only to raise money but also to alter user behavior in order to improve the performance of a transportation system with respect to indicators such as travel time or pollution. In this regard, let $\mathbf{v}_{\mathrm{SO}}$ be a globally optimal solution to the mathematical program

$$\underset{\mathbf{v}\in F}{\text{minimize}}\ \mathbf{t}(\mathbf{v})^{\top}\mathbf{v}, \tag{72}$$

and let us first consider the problem of setting tolls that induce system-optimal flows $\mathbf{v}_{\mathrm{SO}}$, while being compatible with the selfish behavior of users.[24] Then

[23] The reader may draw a piece-wise linear convex function in $\Re$ and see for herself that if there exists a global minimum, then there exists also a 'nondifferentiable' global minimum.

[24] This is a particular instance of (43) with $\mathbf{v}^* = \mathbf{v}_{\mathrm{SO}}$. For notational simplicity, we have adopted a link flow formulation of the equilibrium problem.

any toll vector $\boldsymbol{\rho}$ that satisfies

$$-\big[\mathbf{t}(\mathbf{v}_{\mathrm{SO}}) + \boldsymbol{\rho}\big] \in N_F(\mathbf{v}_{\mathrm{SO}}) \tag{73}$$

clearly satisfies our requirements. We say that Equation (73) describes an *inverse optimization problem* where, given an optimal solution to a mathematical program, one looks for a vector of parameters ($\boldsymbol{\rho}$ in the above) that reproduces the given solution. To pursue the analysis, we assume that the polyhedron F takes the form $F = \{\mathbf{v} \mid \mathbf{B}\mathbf{v} = \mathbf{b}, \mathbf{v} \geqslant \mathbf{0}\}$, so that one can rewrite the toll Equation (73) in the Karush–Kuhn–Tucker format

$$\mathbf{t}(\mathbf{v}_{\mathrm{SO}}) + \boldsymbol{\rho} - \mathbf{B}^{\top}\boldsymbol{\pi} \geqslant \mathbf{0}, \tag{74a}$$

$$\mathbf{v}_{\mathrm{SO}}^{\top}\big[\mathbf{t}(\mathbf{v}_{\mathrm{SO}}) + \boldsymbol{\rho} - \mathbf{B}^{\top}\boldsymbol{\pi}\big] = \mathbf{0}, \tag{74b}$$

$$\mathbf{v}_{\mathrm{SO}} \in F. \tag{74c}$$

Because $\mathbf{v}_{\mathrm{SO}}$ is known and feasible, we obtain that the solution set of (74) associated with the system-optimal flow $\mathbf{v}_{\mathrm{SO}}$ is the polyhedron P defined as

$$\boldsymbol{\rho} \geqslant \mathbf{B}^{\top}\boldsymbol{\pi} - t(\mathbf{v}_{\mathrm{SO}}),$$
$$\boldsymbol{\rho}_i = \big[\mathbf{B}^{\top}\boldsymbol{\pi} - t(\mathbf{v}_{\mathrm{SO}})\big]_i, \quad \text{if } (\mathbf{v}_{\mathrm{SO}})_i > 0.$$

Since $\mathbf{v}_{\mathrm{SO}}$ is solution to the variational inequality

$$-\big[\mathbf{t}(\mathbf{v}_{\mathrm{SO}}) + \mathbf{t}'(\mathbf{v}_{\mathrm{SO}})\mathbf{v}_{\mathrm{SO}}\big] \in N_F(\mathbf{v}), \tag{75}$$

the marginal cost solution $\boldsymbol{\rho} = \mathbf{t}'(\mathbf{v}_{\mathrm{SO}})\mathbf{v}_{\mathrm{SO}}$ is an element of P. If the system-optimal solution is not unique in link-flow space, then the set P must be enlarged to include all polyhedra associated with system-optimal flows. The resulting union of polyhedra is, in the general situation, not convex.

Under suitable assumptions, $\mathbf{v}_{\mathrm{SO}}$ is uniquely defined. However, the set P contains several elements, which suggests optimizing over a secondary criterion. For technical or political reasons, one may wish to minimize the total revenue levied, and insist that link tolls be nonnegative. This yields the mathematical program

$$\underset{\boldsymbol{\rho}\geqslant\mathbf{0},\boldsymbol{\pi}}{\text{minimize}}\ \boldsymbol{\rho}^{\top}\mathbf{v}_{\mathrm{SO}}, \tag{76a}$$

$$\text{subject to}\quad \mathbf{t}(\mathbf{v}_{\mathrm{SO}}) + \boldsymbol{\rho} - \mathbf{B}^{\top}\boldsymbol{\pi} \geqslant \mathbf{0}, \tag{76b}$$

$$\mathbf{v}_{\mathrm{SO}}^{\top}\big(\mathbf{t}(\mathbf{v}_{\mathrm{SO}}) + \boldsymbol{\rho} - \mathbf{B}^{\top}\boldsymbol{\pi}\big) = 0. \tag{76c}$$

Since both terms of the complementarity equation are nonnegative, this constraint can be replaced by the constraint (76c). Writing its dual linear program as that to

$$\underset{\alpha\geqslant 0,\mathbf{v}}{\text{maximize}}\ \mathbf{t}(\mathbf{v}_{\mathrm{SO}})^{\top}(\alpha\mathbf{v}_{\mathrm{SO}} - \mathbf{v}),$$
$$\text{subject to}\quad \mathbf{v} - \alpha\mathbf{v}_{\mathrm{SO}} \leqslant \mathbf{v}_{\mathrm{SO}},$$
$$\mathbf{B}\mathbf{v} = \alpha\mathbf{b},$$

and performing the change of variable $\mathbf{v} = \alpha\mathbf{w}$ yields the parametric multicommodity flow problem to

$$\begin{aligned} &\underset{\alpha\geqslant 0,\mathbf{w}\in F}{\text{maximize}}\ \mathbf{t}(\mathbf{v}_{\text{SO}})^{\top}(\mathbf{v}_{\text{SO}}-\mathbf{w}), \\ &\quad \text{subject to}\quad \mathbf{w}\leqslant\left(1+\frac{1}{\alpha}\right)\mathbf{v}_{\text{SO}}, \end{aligned} \tag{77}$$

which is equivalent to maximize$_{\alpha\geqslant 0}\,\phi(\alpha)$, where

$$\begin{aligned} &\phi(\alpha) := \underset{\mathbf{w}\in F}{\text{maximum}}\ \mathbf{t}(\mathbf{v}_{\text{SO}})^{\top}(\mathbf{v}_{\text{SO}}-\mathbf{w}), \\ &\quad \text{subject to}\quad \mathbf{w}\leqslant\left(1+\frac{1}{\alpha}\right)\mathbf{v}_{\text{SO}}. \end{aligned} \tag{78}$$

The scalar function ϕ has interesting properties: $\phi(0) = 0$, $\phi(\alpha)$ is nonnegative ($\mathbf{v}_{\text{SO}}$ is a feasible solution), ϕ is piecewise linear, ϕ is bounded from above (the primal problem (76) is feasible and bounded), ϕ is concave, and ϕ is increasing.

The properties of the function ϕ imply that there exists a threshold value α^* such that $\phi(\alpha) = \phi(\alpha^*)$ whenever $\alpha \geqslant \alpha^*$. In other words, whenever α is sufficiently large, a *single* multicommodity flow problem needs to be solved. The optimal tolls are then retrieved from the dual vector corresponding to the flow constraints $\mathbf{Bw} = \mathbf{b}$.

In a multi-attribute context, marginal cost pricing may not be applicable any more, as the resulting tolls will be different for distinct classes. It is therefore surprising that 'system-optimal' tolls that achieve the same goal do actually exist. Let us first analyze a two-attribute (time and money) problem where we look for a link toll vector that induces a system-optimal solution with respect to time.[25] As in Section 3.1, we denote by α_g the VOT parameter associated with the link-flow vector $\mathbf{v}^g$ of group $g \in \mathcal{G}$. Let

$$\mathbf{v}_{\text{SO}} \in \arg\underset{\mathbf{v}\in F}{\text{minimum}}\ \mathbf{t}(\bar{\mathbf{v}})^{\top}\bar{\mathbf{v}} + \sum_{g\in\mathcal{G}}\alpha_g\mathbf{f}(\bar{\mathbf{v}})^{\top}\mathbf{v}^g$$

denote the system-optimal solution, and $\bar{\mathbf{v}}_{\text{SO}}$ the associated total link-flow vector. Consider the linear program

$$\begin{aligned} &\underset{\mathbf{v}\in F}{\text{minimize}}\ \mathbf{t}(\bar{\mathbf{v}}_{\text{SO}})^{\top}\bar{\mathbf{v}} + \sum_{g\in\mathcal{G}}\alpha_g\mathbf{f}(\bar{\mathbf{v}}_{\text{SO}})^{\top}\mathbf{v}^g, \\ &\quad \text{subject to}\quad \sum_{g\in\mathcal{G}}\mathbf{v}^g = \bar{\mathbf{v}}_{\text{SO}}, \end{aligned}$$

[25] There also exist system-optimal solutions with respect to cost. One must be aware that they need not coincide.

and the optimal dual vector $\boldsymbol{\alpha}^*$ associated with its sole explicit constraint. Since its primal solution is also a solution to the Lagrangian problem

$$\begin{aligned}&\underset{\mathbf{v}\in F}{\text{minimize}}\, \mathbf{t}(\bar{\mathbf{v}}_{\mathrm{SO}})^\top \bar{\mathbf{v}} + \sum_{g\in\mathcal{G}} \alpha_g \mathbf{f}(\bar{\mathbf{v}}_{\mathrm{SO}})^\top \mathbf{v}^g + \sum_{g\in\mathcal{G}} \alpha_g^* \mathbf{v}^g \\ &\quad = \sum_{g\in\mathcal{G}} \left[\mathbf{t}(\bar{\mathbf{v}}_{\mathrm{SO}}) + \alpha_g \mathbf{f}(\bar{\mathbf{v}}_{\mathrm{SO}}) + \boldsymbol{\alpha}^*\right]^\top \mathbf{v}^g,\end{aligned}$$

it satisfies the Wardrop equilibrium principle for every individual class $g \in \mathcal{G}$. It follows that the dual vector $\boldsymbol{\alpha}^*$ induces system-optimal flows and is a suitable choice for a toll vector. If one insists that tolls be positive, then one can replace the compatibility constraint $\sum_{g\in\mathcal{G}} \mathbf{v}^g = \bar{\mathbf{v}}_{\mathrm{SO}}$ by the inequality $\sum_{g\in\mathcal{G}} \mathbf{v}^g \geqslant \bar{\mathbf{v}}_{\mathrm{SO}}$, which must be tight at the optimal solution to the linear program.

The analysis can be repeated in the infinite-dimensional case, setting tolls to the dual vector of the infinite-dimensional LP to

$$\begin{aligned}&\underset{\mathbf{v}\in F}{\text{minimize}}\, \mathbf{t}(\bar{\mathbf{v}}_{\mathrm{SO}})^\top \bar{\mathbf{v}} + \int_0^\infty \alpha f(\bar{\mathbf{v}}_{\mathrm{SO}})^\top \mathbf{v}(\alpha)\, \mathrm{d}\alpha, \\ &\quad \text{subject to} \quad \int_0^\infty \mathbf{v}(\alpha)\, \mathrm{d}\alpha = \bar{\mathbf{v}}_{\mathrm{SO}}.\end{aligned}$$

By formulating the multi-attribute problem with respect to total flows (see Section 3.1), it is also possible to derive a finite-dimensional formulation variant, and obtain the same result via arguments invoked in the discrete case.

Note that, in the previous analysis, no monotonicity condition was required from the mappings $\mathbf{t}$ and $\mathbf{f}$, and that cycle-free flow patterns other than system-optimal ones could be induced by the above technique. Note also that, if tolls cannot be set on all links of the transportation network, it might not be possible to achieve system optimality through a toll schedule. The resulting problem, which involves equilibrium constraints, can be formulated as an MPEC. This class of hard nonconvex and nondifferentiable problems is usually not amenable to theoretically efficient algorithms, although much attention has been paid, recently, to the design of locally convergent algorithms. Instances of such problems are presented in the next two sections.

6.3 Tolls for raising revenues

Let us reconsider the Braess paradox network illustrated in Figure 4. By imposing a toll ρ_{23} larger than 13 on the diagonal link (2, 3), one can induce the link-flow system-optimal solution

$$\mathbf{v} = (v_{12}, v_{13}, v_{23}, v_{24}, v_{34}) = (3, 3, 0, 3, 3),$$

with equilibrium cost 83. Alternatively, one may wish to maximize the revenue generated by the toll link (2, 3). By solving the equilibrium problem for ρ_{23} fixed, one can express the diagonal flow v_{23} as a function of ρ_{23} and obtain the

revenue

$$\rho_{23} \cdot v_{23}(\rho_{23}) = \begin{cases} 4\rho_{23}, & \text{if } \rho_{23} \leqslant -26, \\ \rho_{23}\big(2 - \frac{2}{13}\rho_{23}\big), & \text{if } \rho_{23} \in [-26, 13], \\ 0, & \text{if } \rho_{23} \geqslant 13. \end{cases}$$

The optimal revenue of 13/2 is achieved when $\rho_{23} = 13/2$. This toll not only maximizes revenue but, surprisingly, also reduces each user perceived travel cost from 92 to 87.5.

In the case of a general network, let $\mathcal{L}_1$ denote the set of toll links and $\mathcal{L}_2 = \mathcal{L} \setminus \mathcal{L}_1$; let $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$, $\boldsymbol{\rho} = (\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) = (\boldsymbol{\rho}_1, \mathbf{0})$, $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$, and $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2)$ be the respective partitions of the link flow vector, toll vector, constraint matrix and cost mapping. The revenue maximizing problem is to

$$\begin{aligned} &\underset{\mathbf{v}\in F, \boldsymbol{\rho}_1}{\text{maximize}}\ \boldsymbol{\rho}_1^\top \mathbf{v}_1, \\ &\quad \text{subject to} \quad \big(\mathbf{t}(\mathbf{v}) + \boldsymbol{\rho}\big)^\top (\mathbf{v} - \mathbf{w}) \leqslant 0, \quad \mathbf{w} \in F, \\ &\qquad\qquad\ \ \boldsymbol{\rho}_2 = \mathbf{0}. \end{aligned} \tag{79}$$

If the mapping $\mathbf{t}$ is monotone, one can replace the lower level variational inequality by its Karush–Kuhn–Tucker conditions. This yields the MPEC formulation

$$\begin{aligned} &\underset{\mathbf{v}\in F, \boldsymbol{\rho}_1, \boldsymbol{\pi}}{\text{maximize}}\ \boldsymbol{\rho}_1^\top \mathbf{v}_1, \\ &\quad \text{subject to} \quad \mathbf{t}_1(\mathbf{v}) + \boldsymbol{\rho}_1 - \mathbf{B}_1^\top \boldsymbol{\pi} \geqslant \mathbf{0}, \\ &\qquad\qquad\ \ \mathbf{t}_2(\mathbf{v}) + \boldsymbol{\rho}_2 - \mathbf{B}_1^\top \boldsymbol{\pi} \geqslant \mathbf{0}, \\ &\qquad\qquad\ \ \mathbf{v}_1^\top \big[\mathbf{t}_1(\mathbf{v}) + \boldsymbol{\rho}_1 - \mathbf{B}_2^\top \boldsymbol{\pi}\big] = 0, \\ &\qquad\qquad\ \ \mathbf{v}_2^\top \big[\mathbf{t}_2(\mathbf{v}) - \mathbf{B}_2^\top \boldsymbol{\pi}\big] = 0, \end{aligned} \tag{80}$$

or, upon the introduction of a vector $\mathbf{z}$ of binary variables and a suitably large constant M, the mixed discrete-continuous program

$$\begin{aligned} &\underset{\mathbf{v}\in F, \boldsymbol{\rho}_1, \boldsymbol{\pi}, \mathbf{z}}{\text{maximize}}\ \boldsymbol{\rho}_1^\top \mathbf{v}_1, \\ &\quad \text{subject to} \quad \mathbf{t}(\mathbf{v}) + \boldsymbol{\rho} - \mathbf{B}^\top \boldsymbol{\pi} \geqslant \mathbf{0}, \\ &\qquad\qquad\ \ \mathbf{v} \leqslant M\mathbf{z} \geqslant \mathbf{t}(\mathbf{v}) + \boldsymbol{\rho} - \mathbf{B}^\top \boldsymbol{\pi}, \\ &\qquad\qquad\ \ z_l \in \{0, 1\},\ l \in \mathcal{L}. \end{aligned} \tag{81}$$

If the mapping $\mathbf{t}$ is monotone and separable, (81) can be approximated by a piecewise linear function, hence the importance of investigating the linear case, where $\mathbf{t}(\mathbf{v}) = \mathbf{t}^\top \mathbf{v}$. First, due the complementarity conditions, $\rho_{1l} = (\mathbf{B}_1^\top \boldsymbol{\pi})_l - t_{1l}$ whenever the flow v_{1l} is positive. On the other hand, ρ_{1l} may be set to any value larger than $(\mathbf{B}_2^\top \boldsymbol{\pi})_l - t_{1l}$ when $v_{1l} = 0$. Therefore, one may set a priori $\rho_1 = \mathbf{B}_1^\top \boldsymbol{\pi}$ in the optimal solution. Replacing ρ_1 by its value in the objective

of (80) one obtains the simplified formulation

$$\begin{aligned} &\underset{\mathbf{v}\in F,\boldsymbol{\pi}}{\text{maximize}} \left(\mathbf{B}_1^\top \boldsymbol{\pi} - \mathbf{t}_1\right)^\top \mathbf{v}_1 = \boldsymbol{\pi}^\top (\mathbf{b} - \mathbf{B}_2 \mathbf{v}_2) - \mathbf{t}_1^\top \mathbf{v}_1 = \mathbf{b}^\top \boldsymbol{\pi} - \mathbf{t}^\top \mathbf{v}, \\ &\quad \text{subject to} \quad \mathbf{t}_2 - \mathbf{B}_2^\top \boldsymbol{\pi} \geqslant 0, \\ &\qquad\qquad\quad \mathbf{v}_2^\top \left(\mathbf{t}_2 - \mathbf{B}_2^\top \boldsymbol{\pi}\right) = 0. \end{aligned} \tag{82}$$

If one solves a mixed integer reformulation of (82) by an implicit enumeration method, an upper bound on the revenue must be available at each node of the branch-and-bound tree. At such a node, let $\mathcal{L}_0 \subseteq \mathcal{L}_2$ denote the index set of toll-free links with null flow, and let us relax the complementarity constraint corresponding to flows $\mathbf{v}_l$, $l \in \mathcal{L}_2 \setminus \mathcal{L}_0$, yielding the upper bound

$$\begin{aligned} &\underset{\boldsymbol{\pi}}{\text{maximum}}\, \mathbf{b}^\top \boldsymbol{\pi} \qquad - \qquad \underset{\mathbf{v}\in F}{\text{minimum}}\, \mathbf{t}^\top \mathbf{v}, \\ &\quad \text{subject to} \quad B_2^\top \boldsymbol{\pi} \leqslant \mathbf{t}_2, \\ &\qquad\qquad\quad \left(B_2^\top \boldsymbol{\pi}\right)_l = t_{2l}, \quad l \in \mathcal{L}_0. \end{aligned}$$

Replacing the left-hand side LP by its dual, the upper bound can be expressed as

$$\begin{aligned} &\underset{\mathbf{v}\in F,\mathbf{v}_1=\mathbf{0}}{\text{minimum}}\, \mathbf{t}^\top \mathbf{v} \qquad - \qquad \underset{\mathbf{v}\in F}{\text{minimum}}\, \mathbf{t}^\top \mathbf{v}, \\ &\quad \text{subject to} \quad v_{2l} \text{ free if } l \in \mathcal{L}_0, \end{aligned} \tag{83}$$

and can be interpreted as the difference between two shortest distances, one in a network with tolls set at 0 and some link flows unrestricted in sign, the other in a network with tolls set at ∞. Once the subset of positive flows has been selected, a set of optimal tolls compatible with this choice can be retrieved from the dual solution associated with the LP that appears on the left-hand side of (83). This operation, which amounts to solving an *inverse* linear program, can be performed very efficiently, by computing shortest paths in a network from which toll links have been removed, and where backward copies of toll-free links carrying positive flow are introduced.

6.4 A continuous network design problem

In this section, we address the problem that consists in designing a network subject to equilibrium constraints, through the specification of continuous control parameters. More specifically, let $\varphi(\boldsymbol{\rho})$ denote the cost of implementing a traffic control $\boldsymbol{\rho}$ belonging to some set Z, assumed to be convex. Based on this notation, the toll problem considered in Section 6.2 constitutes special instances of the general problem to

$$\begin{aligned} &\underset{\boldsymbol{\rho}\in Z,\mathbf{v}\in F(\boldsymbol{\rho})}{\text{minimize}}\, \mathbf{v}^\top \mathbf{t}(\mathbf{v},\boldsymbol{\rho}) + \varphi(\boldsymbol{\rho}), \\ &\quad \text{subject to} \quad -\mathbf{t}(\mathbf{v},\boldsymbol{\rho}) \in N_F(\boldsymbol{\rho})(\mathbf{v}), \end{aligned} \tag{84}$$

where the vector of design parameters $\boldsymbol{\rho}$ may impact both the cost mapping $\mathbf{t}$ and the feasible set $F(\boldsymbol{\rho})$. For instance, one may consider improving a road network through capacity enhancement, balancing long term investment costs against recurrent travel delays. In this context, let φ_l represent the capacity of link l, $\varphi_l(z_l)$ the cost of achieving capacity z_l, where we assume that the functions $\mathbf{t}$ and φ are link-separable. If, furthermore, we assume that t_l is a function of the flow-capacity ratio, that the set F is independent of $\boldsymbol{\rho}$, and that the design vector $\boldsymbol{\rho}$ is unconstrained, then a continuous variant of the network design problem (84) takes the form

$$\begin{aligned} &\underset{\boldsymbol{\rho}\in Z,\mathbf{v}\in F}{\text{minimize}} \sum_{l\in\mathcal{L}} v_l t_l\left(\frac{v_l}{\rho_l}\right) + \varphi_l(\rho_l), \\ &\quad \text{subject to} \quad \mathbf{v} \in \arg\underset{\mathbf{w}\in F}{\text{minimum}} \sum_{l\in\mathcal{L}} \int_0^{w_l} t_l\left(\frac{u}{\rho_l}\right) \mathrm{d}u. \end{aligned} \tag{85}$$

While the resulting problem is theoretically difficult, it is amenable to efficient heuristic procedures. Maybe the simplest one consists in first solving the system-optimal problem to

$$\text{H1:} \quad \underset{\boldsymbol{\rho}\in Z,\mathbf{v}\in F}{\text{minimize}} \sum_{l\in\mathcal{L}} v_l t_l\left(\frac{v_l}{\rho_l}\right) + \varphi_l(\rho_l), \tag{86}$$

for the design vector $\boldsymbol{\rho}_{\mathrm{SO}}$ and then finding the equilibrium flows compatible with $\boldsymbol{\rho}_{\mathrm{SO}}$. This is Heuristic H1. Another natural procedure consists in iterating, à la Gauss–Seidel, between flow and capacity assignment subproblems: for fixed $\boldsymbol{\rho}$, one solves a traffic equilibrium problem while, for fixed flow pattern $\mathbf{v}$, each $\rho_l = \rho_l(v_l)$ satisfies the single-variable equation

$$\text{H2:} \quad t_l\left(\frac{v_l}{\rho_l}\right) + \left(\frac{v_l}{\rho_l}\right)^2 t_l'\left(\frac{v_l}{\rho_l}\right) + \varphi_l'(\rho_l) = 0, \tag{87}$$

and is frequently available in closed form. If the procedure converges, it does so to an equilibrium flow that is compatible with the 'greedy' optimality condition (87). This is Heuristic H2, whose solution can also be computed by solving the single variational inequality

$$-\mathbf{t}\big(\mathbf{v}, \boldsymbol{\rho}(\mathbf{v})\big) \in N_F(\mathbf{v}). \tag{88}$$

A general class of heuristic procedures consists in solving a problem where the flow part of the solution automatically obeys Wardrop's equilibrium principle. This can be achieved by replacing the total cost by its integral in the objective and, for generality, scaling the investment cost term. This yields the mathematical program

$$\text{H3:} \quad \underset{\boldsymbol{\rho}\in Z,\mathbf{v}\in F}{\text{minimize}}\, \mathbf{t}(\mathbf{v}, \boldsymbol{\rho}) = \sum_{l\in\mathcal{L}} \int_0^{v_l} t_l\left(\frac{u}{\rho_l}\right) \mathrm{d}u + \sum_{l\in\mathcal{L}} \xi_l \varphi_l(\rho_l). \tag{89}$$

Next, one might look for the set of parameters that yields the best solution to the original program (85). This problem, however, is of the same computational complexity as solving (85) directly, and is thus 'intractable'. In this respect, it is natural to restrict our attention to a subfamily H3, for instance that where all ξ_l are set to a common value ξ. To pursue the analysis, we suppose that φ is the monomial $\varphi_l(\rho_l) = d_l\rho_l^m$ and that the cost function $\mathbf{t}$ assumes the BPR form $t_l(x) = \alpha_l + \beta_l x^p$. Under these assumptions, the solution to (87) is given by

$$\rho_l(v_l) = \left(\frac{p\beta_l}{md_l}\right)^{1/(p+m)} v_l^{(p+1)/(p+m)}. \tag{90}$$

If one introduces the function

$$\zeta_l(v_l) = \left(\frac{1}{p+1}\right)^{1/p} \left(\frac{p\beta_l}{md_l}\right)^{1/(p+m)} v_l^{(p+1)/(p+m)},$$

one may check that the solution to the convex program

$$\text{H4:} \quad \underset{\mathbf{v}\in F}{\text{minimize}} \sum_{l\in\mathcal{L}} \int_0^{v_l} \left[\alpha_l + \beta_l\left(\frac{u}{\rho_l(u)}\right)^p\right] \mathrm{d}u \tag{91}$$

coincides with the system-optimal flow of program (86).

Note that, for the above functional forms, the class H3 defined by (89) subsumes the Gauss–Seidel solution, the latter corresponding to the choice $\xi_l = 1/(p+1)$ for every $l \in \mathcal{L}$, and that all heuristics presented find a design vector z by solving a convex optimization problem, and are thus easy to implement. To assess the quality of their respective solutions, we adopt the worst-case analysis point of view, and aim to determine (or estimate) the ratio

$$\begin{aligned} R_m^p(H) &= \sup_{\alpha,\beta,d} \frac{\text{cost of heuristic solution}}{\text{cost of (unknown) optimal solution}} \\ &= \sup_{\alpha,\beta,\Delta} \frac{F(\mathbf{v}(\boldsymbol{\rho}^{\boldsymbol{H}}), \boldsymbol{\rho}^H)}{F(\mathbf{v}^*, \boldsymbol{\rho}^*)} \end{aligned} \tag{92}$$

over all possible network configurations represented by the symbol Δ. This analysis is reminiscent of the worst-case analysis presented in Section 2.9. We have the following approximation results.

Theorem 21. *For heuristics* H1 *through* H4*, the following bounds on* R_m^p *hold:*

(a) $\lim_{p\to\infty} R_1^p(\text{H1}) \geqslant 2$;

(b) $R_1^p(\text{H2}) = p + 1$;

(c) $R_m^p(\text{H4}) = m(p+1)/(m+p) + p(m+p)(p+1)^{-m/p}$;

(d) $\lim_{m\downarrow 0} R_m^p(\text{H4}) = 1$ *and* $\lim_{p\to\infty} R_1^p(\text{H4}) = 2$;

(e) $1 + \frac{p}{\xi(p+1)} \leqslant R_1^p(\text{H3}) \leqslant \frac{\xi^{p/(p+1)}}{(p+1)^{1/(p+1)}}[1 + \frac{p}{\xi(p+1)}]^2$;

(f) $2 \leqslant \lim_{p\to\infty} R_1^p(\mathrm{H4}) = 4$.

Remark 22. (i) For linear investment functions ($m = 1$), the value that yields the tightest upper bound on $R_1^p(\mathrm{H3})$ is $\xi = 2 - p/(p+1)$.

(ii) The bounds provided by the theorem are very pessimistic. Indeed, numerical experiments show that the error is of the order of one percent. This is not surprising if one observes that the objective of both 'players' are not antagonistic: the leader actually wishes to minimize the users' travel times.

(iii) If the investment function φ is linear, Equation (87) fixes a value for the ratio v_l/ρ_l. It follows that $\mathbf{t}(\mathbf{v}, \boldsymbol{\rho}(\mathbf{v}))$ is constant in the variational inequality (88) and that its solution is an extremal flow pattern that can be efficiently computed by shortest routes methods.

Once a heuristic solution is obtained, it can be improved using the sensitivity results of Section 6.1 with respect to the design (toll) vector $\boldsymbol{\rho}$. At differentiable points, the procedure reduces to gradient projection. At points of nondifferentiability, a descent direction can be tentatively built from the coordinate-wise directional derivatives. In the numerical example below, these were always sufficient to provide a descent direction at every iteration.

We consider the special case of problem (85) where Z is a box, declaring that $0 \leqslant \rho_l \leqslant u_l$ must hold for every l. An implementation of the DSD algorithm, the directional derivative problem, and a gradient projection algorithm with a rather simple line search,[26] form a complete descent-based framework for this network design problem. We have applied the procedure to a problem defined for the Sioux Falls network (with 10 potential links to improve), and compared it to some heuristic schemes that have been proposed for the continuous network design problem. The results are shown in Table 8.

In the table, we use the following short-hand notation: SBD stands for 'sensitivity-based descent', that is, the algorithm presented in this section; H–J stands for 'Hooke–Jeeves', a direct search method based only on objective values; EDO stands for 'equilibrium decomposed optimization', an algorithm which utilizes the 'classical' sensitivity analysis formula (71); SA stands for 'simulated annealing'; and PIPA stands for 'penalty interior point algorithm'. The Notes section provides the sources for the results in the table.

The values in the last row of the table corresponds to a much more accurately calculated upper-level objective value for the terminal designs; in the case of our algorithm, the difference between the nominal values are small, since accurate traffic equilibrium computations are used throughout. For the other algorithms – which typically use 50–100 Frank–Wolfe iterations on each lower-level problem – the differences are much larger. Clearly, their final solutions suffer from a poor estimation of traffic equilibrium solutions, in addition

[26] We have used the Armijo step length rule; the main motivation is that the objective value calculations are quite expensive.

Table 8.
Results for network design on the Sioux Falls network

	SBD	SBD	SBD	SBD	H–J	EDO	SA	H–J	EDO	SA	PIPA
Initial value $\boldsymbol{\rho}^0$:	0	2	3	6.25	2	–	6.25	3	0	6.25	–
Upper bound $\mathbf{u}$:	25	25	25	25	–	25	25	–	25	25	–
ρ_{16}	5.3027	5.1492	5.3457	5.2773	4.8	4.59	5.38	4.507	4.276	5.322	5.4680
ρ_{17}	2.0560	2.0214	1.9786	2.0533	1.2	1.52	2.26	4.509	2.288	2.596	2.0039
ρ_{19}	5.3430	5.1679	5.3741	5.3002	4.8	5.45	5.50	4.520	4.080	5.664	5.4471
ρ_{20}	1.9901	2.0012	1.9460	2.0369	0.8	2.33	2.01	4.052	1.618	1.309	1.9395
ρ_{25}	2.5216	2.4945	2.7856	2.7670	2.0	1.27	2.64	4.299	1.654	2.498	2.9448
ρ_{26}	2.5548	2.5447	2.8245	2.8222	2.6	2.33	2.47	2.949	1.130	2.732	2.8191
ρ_{29}	2.9883	2.9535	2.9257	3.0124	4.8	0.41	4.54	3.000	3.219	4.123	3.4039
ρ_{39}	4.8559	4.8330	4.7528	4.7348	4.4	4.59	4.45	3.601	3.326	4.508	4.8061
ρ_{48}	3.0026	2.9798	2.9732	2.9746	4.8	2.71	4.21	3.006	1.981	3.736	3.2364
ρ_{74}	4.8496	4.8212	4.7347	4.7511	4.4	2.71	4.67	3.200	3.190	3.903	4.7779
Obj, reported	79.9969	79.9968	79.9987	80.0026	80.78	83.08	80.87	83.316	83.703	81.983	80.8669
Obj, recomputed	79.9961	79.9971	79.9990	80.0043	80.67	82.34	80.42	81.185	81.345	80.304	80.0528

to being further from a local optimum (or, a stationary point) than in the proposed algorithm.

The reason why traffic equilibrium computations need to be accurate becomes obvious when one realizes that they serve as input to the sensitivity analysis procedure, which in turn provides the gradient values or, at least, coordinate-wise directional derivatives. Unless the equilibrium solution is accurately computed, the sensitivity analysis may provide erroneous results, leading to bad search directions, or to the algorithm halting prematurely. However, this is not a serious problem with the DSD algorithm, which can solve the traffic equilibrium problem to a sufficiently good accuracy quickly. This must be taken as an indication that the Frank–Wolfe method, or any other slowly converging method for the traffic equilibrium problem, should definitely *not* be used.

We remark finally that the necessity to provide accurate equilibrium solutions arises in several other applications involving traffic equilibrium models. We mention here only two, in addition to all models of a hierarchical nature, such as traffic management or toll optimization type models: (i) assessing the environmental impact of the transportation network (exhaust fumes, for instance) relies on accurate link usage data, which in turn rely on accurate traffic equilibrium solutions; (ii) travel forecasting, subject to changes in the network configuration, also relies on accurate flow estimates, both for the 'before' and 'after' simulations.

6.5 Bibliographical notes

The sensitivity analysis of Section 6.1 was first developed in Patriksson and Rockafellar (2002, 2003) and later refined in Patriksson (2004). (Regarding

the background material, (64) stems from Dontchev and Rockafellar, 2001, and (66) from Kyparisis, 1990; see also Robinson 1980, 1985, 1991.) The material presented here is a condensed version that for the main part discusses the standard, separable model. That the sensitivity analysis does not depend on the equilibrium route flow chosen was first noticed in Tobin and Friesz (1988), albeit in a framework that has limitations, as we remark and show; this independence was observed and utilized later in Patriksson and Rockafellar (2002, 2003), Patriksson (2004), Patriksson (2006). The examples showing the limitations of the Tobin/Friesz formula (71) are taken from Patriksson and Josefsson (2003), Josefsson and Patriksson (2005), Patriksson (2006).

While it is a classical result from economic theory that marginal cost pricing can induce a system-optimal flow pattern, Section 6.2 focuses on toll schedules that achieve that aim at minimal user cost. Since the set of system-optimal tolls is polyhedral (see Bergendorff et al., 1997; Hearn and Ramana, 1998; Larsson and Patriksson, 1998), such schedules are solutions to a linear program. Dial (1999, 2000), either in the single or multicommodity case, proposed an algorithm that exploits the underlying network structure. A similar result was obtained by Marcotte and Savard (2002), who made the link, in the single-commodity case, with the inverse shortest route problem (see Ahuja and Orlin, 2001). In a multi-attribute context, where marginal taxation may not be feasible, Yang and Huang (2004) gave a constructive proof of the existence of system-optimal tolls. The extension of their method from the discrete to the continuous case is new to this book. A more complex existence proof, of a highly nonconstructive nature (even for the discrete case), was given by Cole et al. (2003).

Whenever tolls cannot be set on *every* link of a network, the nature of the problem becomes more complex, and falls within the realm of *bilevel programming*. Although bilevel programs, together with the closely related MPECs (Mathematical Programs with Equilibrium Constraints), have been topics of several studies in the past years, much work remains to be done before robust procedures may address real-life instances. The reader is referred to the books by Luo et al. (1996), Bard (1998), and Shimizu et al. (1997) for introductions to the subject.

Section 6.3 addresses one such instance, where one aims at maximizing toll revenues. References for theory, algorithms, and applications relevant to uncongested networks are Labbé et al. (1998, 1999), Brotcorne et al. (2001), Marcotte and Savard (2001, 2002), Bouhtou et al. (2003), Côté et al. (2003). A linearization algorithm applicable to the congested case can be found in Julsain (1998).

The continuous variant of the network design problem considered in Section 6.4 was first proposed in Abdulaal and Leblanc (1979). Heuristics were analyzed from the theoretical and computational perspectives in Marcotte (1986b) and Marcotte and Marquis (1992). The numerical results displayed in Figure 10 and Table 8 demonstrate that Frank–Wolfe type methods are to recommend for use neither in the solution of the basic traffic equilibrium

model nor for more complex models where accurate equilibrium solutions are needed as inputs. The numerical results are taken from Patriksson and Josefsson (2003), Patriksson (2006). In Table 8, columns five and six appeared in Suwansirikul et al. (1987), column seven in Friesz et al. (1992), columns eight–ten in Huang and Bell (1998), and column eleven in Lim (2002). The remarks made at the end of the section refer to application work performed in Larsson et al. (2001), Boyce et al. (2002).

Appendix A: A primer on variational inequalities

A.1 Definition and formulations

Equilibrium problems naturally fit the variational inequality framework, which constitutes an extension of equation systems, in contrast with optimization problems. A *variational inequality problem* VIP($\mathbf{f}, X$) in $\Re^n$ is characterized by a nonempty, closed and convex set $X \subseteq \Re^n$ and a mapping $\mathbf{f}$ from $\Re^n$ into $\Re^n$. A vector $\mathbf{x} \in X$ is a solution of VIP($\mathbf{f}, X$) if and only if it satisfies

$$\mathbf{f}(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \geqslant 0, \quad \mathbf{y} \in X. \tag{93}$$

If one denotes by $N_X(\mathbf{x})$ the normal cone to the set X at the point $\mathbf{x} \in X$, defined as

$$N_X(\mathbf{x}) = \left\{ \mathbf{p} \in \Re^n \middle| \mathbf{p}^\top (\mathbf{y} - \mathbf{x}) \leqslant 0, \quad \mathbf{y} \in X \right\}, \tag{94}$$

then $\mathbf{x}$ is a solution to VIP($\mathbf{f}, X$) if and only if it satisfies the *generalized equation* (or, *normal cone inclusion*)

$$-\mathbf{f}(\mathbf{x}) \in N_X(\mathbf{x}). \tag{95}$$

If $\mathbf{f}$ is a gradient mapping, that is, there exists a real function ϕ such that $\mathbf{f} \equiv \nabla \phi$, then solving VIP($\mathbf{f}, X$) reduces to finding a point that satisfies the first-order optimality condition of the mathematical program

$$\underset{\mathbf{x} \in X}{\text{minimize}}\ \phi(\mathbf{x}). \tag{96}$$

If $X = \{\mathbf{x} \in \Re^n \mid g_i(\mathbf{x}) \leqslant 0, i = 1, \ldots, m\}$ and the functions g_i, $i = 1, \ldots, m$, are continuously differentiable, then, under suitable constraint qualifications (CQs), any solution $\mathbf{x}$ to VIP($\mathbf{f}, X$), together with a vector of multipliers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$, satisfies the Karush–Kuhn–Tucker system

$$g_i(\mathbf{x}) \leqslant \mathbf{0}^m, \tag{97a}$$

$$\mathbf{f}(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}) = \mathbf{0}^n, \tag{97b}$$

$$\lambda_i g_i(x) = 0, \quad i = 1, \ldots, m, \tag{97c}$$

$$\lambda_i \geqslant 0, \quad i = 1, \ldots, m. \tag{97d}$$

If $X = \Re^n_+$, the Karush–Kuhn–Tucker system reduces to the *nonlinear complementarity* problem

$$\mathbf{x} \geqslant \mathbf{0}^n, \quad \mathbf{f}(\mathbf{x}) \geqslant \mathbf{0}^n, \quad \mathbf{x}^\top \mathbf{f}(\mathbf{x}) = 0, \tag{98}$$

or, using a more 'geometric' shorthand notation:

$$\mathbf{0}^n \leqslant \mathbf{x} \perp \mathbf{f}(\mathbf{x}) \geqslant \mathbf{0}^n. \tag{99}$$

If $X \neq \Re^n_+$, the KKT system (97) can still be formulated as a nonlinear complementarity problem, through a suitable redefinition of the primal and dual variables. It follows that both formulations are equivalent.

Variational inequalities can also be formulated as fixed point problems. Two such examples, involving either the projection operator Proj_X over the set X, or a linear program defined over X, are

$$\mathbf{x} = \mathrm{Proj}_X\big(\mathbf{x} - \alpha \mathbf{f}(\mathbf{x})\big), \quad \alpha > 0; \tag{100}$$

$$\mathbf{x} \in \arg \operatorname*{minimum}_{\mathbf{y} \in X} \mathbf{f}(\mathbf{x})^\top (\mathbf{x} - \mathbf{y}). \tag{101}$$

A.2 Monotonicity properties

A mapping $\mathbf{f}$ is *monotone* over X if

$$\big[\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\big]^\top (\mathbf{x} - \mathbf{y}) \geqslant 0, \quad \mathbf{x}, \mathbf{y} \in X. \tag{102}$$

It is *strictly monotone* over X if the above inequality is strict whenever $\mathbf{x} \neq \mathbf{y}$ and *strongly monotone* if there exists a positive number β such that

$$\big[\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\big]^\top (\mathbf{x} - \mathbf{y}) \geqslant \beta \|\mathbf{x} - \mathbf{y}\|^2, \quad \mathbf{x}, \mathbf{y} \in X. \tag{103}$$

If $\mathbf{f}$ is continuously differentiable, it is monotone over X if and only if its Jacobian $\nabla \mathbf{f}(\mathbf{x})$ is positive semidefinite over X, that is, $(\mathbf{y} - \mathbf{x})^\top \nabla \mathbf{f}(\mathbf{x})(\mathbf{y} - \mathbf{x}) \geqslant 0$ holds for every pair $\mathbf{x}, \mathbf{y} \in X$.[27] Monotonicity is to variational inequality problems what convexity is to optimization. Indeed, if $\mathbf{f}$ is the gradient of ϕ, then $\mathbf{f}$ is (strictly, strongly) monotone if and only if ϕ is (strictly, strongly) convex. Other related concepts of interest used in this chapter are:

pseudomonotonicity:

$$\mathbf{f}(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \geqslant 0 \quad \Longrightarrow \quad \mathbf{f}(\mathbf{y})^T (\mathbf{y} - \mathbf{x}) \geqslant 0, \quad \mathbf{x}, \mathbf{y} \in X;$$

monotonicity$^+$:

$$\big[\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\big]^\top (\mathbf{x} - \mathbf{y}) = 0 \quad \Longrightarrow \quad \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{y}).$$

[27] Beware: this is not equivalent to $\nabla \mathbf{f}(\mathbf{x})$ being positive semidefinite for every $\mathbf{x} \in X$.

A mapping that is monotone (respectively pseudomonotone) and satisfies the additional condition that

$$\mathbf{f}(\mathbf{x})^\top(\mathbf{x}-\mathbf{y}) = \mathbf{f}(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}) = 0 \quad \Longrightarrow \quad \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{y})$$

is *monotone*$_*^+$ (respectively *pseudomonotone*$_*^+$). Pseudomonotonicity$_*^+$ ensures that the value of $\mathbf{f}$ is constant over the solution set of VIP($\mathbf{f}$, X). Note that the stronger monotone$_+$ condition is satisfied by convex gradient mappings.

If $\mathbf{f}$ is pseudomonotone and continuous, an equivalent formulation of the variational inequality (93) is the Minty (or dual) variational inequality

$$\mathbf{f}(\mathbf{y})^\top(\mathbf{y}-\mathbf{x}) \geqslant 0, \quad \mathbf{x} \in X. \tag{104}$$

A.3 Existence and uniqueness

Assume that the mapping $\mathbf{f}$ is continuous. If the set X is compact, the solution set X^* of VIP($\mathbf{f}$, X) is nonempty (and compact). If X is closed but unbounded, *coercivity* of $\mathbf{f}$, that is, the existence of a point $\mathbf{x}^0 \in X$ such that

$$\lim_{\|\mathbf{x}\|\to\infty, \mathbf{x}\in X} \frac{\|\mathbf{f}(\mathbf{x})\|}{\|\mathbf{x}-\mathbf{x}^0\|} = \infty,$$

yields the same conclusion. Uniqueness of the solution usually requires that the mapping $\mathbf{f}$ be strictly monotone over X.

A.4 Algorithms

The fixed point formulation (100) suggests the iteration

$$\mathbf{x}^{\tau+1} = \mathrm{Proj}_X\big(\mathbf{x}^\tau - \alpha\mathbf{f}(\mathbf{x}^\tau)\big), \quad \tau = 1, 2, \ldots, \tag{105}$$

which converges if $\mathbf{f}$ is strongly monotone on X and α is sufficiently small, although it is possible to modify this basic scheme and achieve convergence under weaker assumptions. Although similar strategy is not directly applicable to the fixed point formulation (101), it is yet possible to construct an algorithm that finds a solution of VIP($\mathbf{f}$, X) by minimizing the *gap function*

$$\mathrm{gap}(\mathbf{x}) = \operatorname*{maximum}_{\mathbf{y}\in X} \mathbf{f}(\mathbf{x})^\top(\mathbf{x}-\mathbf{y}). \tag{106}$$

While the above function is in general neither convex nor differentiable, one can show that all of its stationary points are solutions to VIP($\mathbf{f}$, X), provided that $\mathbf{f}$ is monotone on X. The descent direction, built around the solutions of the LP problem to

$$\operatorname*{minimize}_{\mathbf{y}\in X} \mathbf{f}(\mathbf{x})^\top\mathbf{y},$$

can, for traffic equilibrium models, be obtained simply by computing shortest route trees rooted at each origin (or destination) node of the network. Alternatively, under pseudomonotonicity, one can minimize the dual gap function

$$\text{dgap}(\mathbf{x}) = \underset{\mathbf{y}\in X}{\text{maximum}}\, \mathbf{f}(\mathbf{y})^{\top}(\mathbf{x}-\mathbf{y}). \tag{107}$$

The resulting convex program can be expressed as a semi-infinite linear program and solved using cutting-plane methods.

A general framework for solving variational inequality problems consists in iteratively solving an approximation of the original problem (93), of a simpler form. One sets $\mathbf{x}^{\tau+1}$ to the solution of VIP($\tilde{\mathbf{f}}(\mathbf{x}^{\tau},\cdot), X$), where $\tilde{\mathbf{f}}$ satisfies the condition that $\tilde{\mathbf{f}}(\mathbf{x},\mathbf{x}) = \mathbf{f}(\mathbf{x})$ for all $\mathbf{x}$ in X. Under relevant choices of the approximation mapping $\tilde{\mathbf{f}}$, one recovers the Jacobi, Gauss–Seidel, Newton, quasi-Newton and projection algorithms, and many others. In particular, if ψ is a strongly convex function, the approximation

$$\tilde{\mathbf{f}}(\mathbf{x},\mathbf{y}) = \mathbf{f}(\mathbf{x}) + \nabla\psi(\mathbf{y}) - \nabla\psi(\mathbf{x}), \quad \mathbf{x},\mathbf{y}\in\Re^n, \tag{108}$$

suggests the continuously differentiable merit function

$$\phi(\mathbf{x}) := \underset{\mathbf{y}\in X}{\text{maximum}}\big[\mathbf{f}(\mathbf{x}) - \nabla\psi(\mathbf{x})\big]^{\top}(\mathbf{x}-\mathbf{y}) - \psi(\mathbf{y}), \quad \mathbf{x}\in X. \tag{109}$$

Under suitable monotonicity conditions on $\mathbf{f}$, the solution $\mathbf{y}(\mathbf{x})$ to (109) yields a feasible descent direction $\mathbf{p} := \mathbf{y}(\mathbf{x}) - \mathbf{x}$ for ϕ, around which a convergent algorithm can be designed. In particular, if one sets $\psi(\mathbf{x}) = \frac{\alpha}{2}\|\mathbf{x}\|^2$, the solution to (109) is precisely the projection $\text{Proj}_X(\mathbf{x} - \alpha\mathbf{f}(\mathbf{x}))$.

A.5 Bibliographical notes

A classical reference on variational inequality problems, mostly in the infinite-dimensional setting, is Glowinski et al. (1981). Variational inequalities were introduced by Hartman and Stampacchia (1966). Several results pertaining to finite-dimensional variational inequalities appear in the monograph Auslender (1976). A more recent and comprehensive treatment is the twin volume Facchinei and Pang (2003a, 2003b). Algorithms based on cost approximations, of which the projection method is the quintessential example, have been analyzed in Cohen (1988), Larsson and Patriksson (1994a), Patriksson (1994a, 1994b, 1998) and several others.

Appendix B: A summary of key notations

$\mathbf{h} = (h_r)$:	vector of flows on routes indexed by r
$\mathbf{c}(\mathbf{h}) = (c_r(\mathbf{h}))$:	vector of route travel costs
$\mathbf{d} = (d_k)$:	vector of demands for commodity (OD pair) k
$\boldsymbol{\pi} = (\pi_{pq})$:	vector of travel costs for origin–destination couples (p, q)
$\mathbf{g}(\boldsymbol{\pi})$:	demand function
$\boldsymbol{\Gamma}$:	route-OD incidence matrix
$\mathbf{v} = (v_l)$:	vector of flows along links indexed by l
$\mathbf{t}(\mathbf{v}) = (t_l(\mathbf{v}))$:	vector of traversal costs on links indexed by l
$\boldsymbol{\Lambda}$:	link–route incidence matrix
$\mathbf{E}$:	link–node incidence matrix
$\mathbf{w}_k$:	vector of OD flows
$\boldsymbol{\xi}(\mathbf{d})$:	inverse demand function: $\boldsymbol{\xi} = \mathbf{g}^{-1}$
H_d:	set of route flows and demand vectors (variable demand case)
H:	set of route flows and demand vectors (fixed demand case)
$\widehat{H}_d$:	set of link flows and demand vectors (link–route representation)
$\widehat{F}$:	set of link flows (link–route representation, fixed demand)
F:	set of link flows (link–node representation, fixed demand)
$\widehat{F}_d$:	set of link flows (link–route representation, variable demand)
F_d:	set of link flows (link–node representation, variable demand)
$\mathbf{f}(\mathbf{v})$:	vector of out-of-pocket travel costs along links
$N_X(\mathbf{x})$:	normal cone to a convex set X at a point $\mathbf{x}$
ϕ_l:	frequency on transit line l
$\boldsymbol{\varphi}(\boldsymbol{\rho})$:	cost vector associated with a link-improvement vector $\boldsymbol{\rho}$
$\mathbf{c}(\mathbf{x})$:	vector of link costs associated with strategic flow vector $\mathbf{x}$
E^s_j:	preference order of strategy s at node j
π^s_{jk}:	probability of accessing node k from node j, when using strategy s

References

Aashtiani, H.Z., Magnanti, T.L. (1981). Equilibria on a congested transportation network. *SIAM Journal on Algebraic and Discrete Methods* 2, 213–226.

Abdulaal, M., Leblanc, L.J. (1979). Continuous equilibrium network design models. *Transportation Research B* 13, 19–32.

Ahuja, R.K., Orlin, J.B. (2001). Inverse optimization. *Operations Research* 49, 771–783.

Ahuja, R.K., Magnanti, T.L., Orlin, J.B. (1993). *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Akamatsu, T. (1996). Cyclic flows, Markov process and stochastic traffic assignment. *Transportation Research B* 30, 369–386.

Altman, E., El Azouzi, R., Abramov, V. (2002). Non-cooperative routing in loss networks. *Performance Evaluation* 49, 43–55.

Auslender, A. (1976). *Optimisation: Méthodes Numériques*. Masson, Paris.

Baillon, J.B. (1975). Un théorème de type ergodique pour les contractions non linéaires dans un espace de Hilbert. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris), Série A* 280, A1511–A1514.

Bard, J.F. (1998). *Practical Bilevel Optimization: Algorithms and Applications*. Kluwer Academic, Boston, MA.

Bar-Gera, H. (2002a). Origin-based algorithm for the traffic assignment problem. *Transportation Science* 36, 398–417.

Bar-Gera, H. (2002b). Origin-based network assignment. In: Patriksson, M., Labbé, M. (Eds.), *Transportation Planning: State of the Art*. Kluwer Academic, Dordrecht, The Netherlands, pp. 1–17.

Beckmann, M., McGuire, C.B., Winsten, C.B. (1956). *Studies in the Economics of Transportation*. Yale Univ. Press, New Haven, CT.

Bennett, L.D. (1993). The existence of equivalent mathematical programs for certain mixed equilibrium traffic assignment problems. *European Journal of Operational Research* 71, 177–187.

Bergendorff, P. (1995). The bounded flow approach to congestion pricing. Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden.

Bergendorff, P., Hearn, D.W., Ramana, M.V. (1997). Congestion toll pricing of traffic networks. In: Pardalos, P.M., Hearn, D.W., Hager, W.W. (Eds.), *Network Optimization*. Proceedings of the Network Optimization Conference, University of Florida, Gainesville, FL, February 12–14, 1996. *Lecture Notes in Economics and Mathematical Systems*, vol. 450. Springer-Verlag, Berlin, pp. 51–71.

Bernstein, D., Smith, T.E. (1994). Equilibria for networks with lower semicontinuous costs: With an application to congestion pricing. *Transportation Science* 28, 221–235.

Bernstein, D., Wynter, L. (2000). Issues of uniqueness and convexity in non-additive bi-criteria equilibrium models. Conference paper, 8th Meeting of the EURO Working Group on Transportation, Rome Jubilee 2000 Conference, La Sapienza, Rome, Italy, September 11–14.

Bertsekas, D.P. (1976). On the Goldstein–Levitin–Polyak gradient projection method. *IEEE Transactions on Automatic Control* 21, 174–184.

Bertsekas, D.P. (1998). *Network Optimization: Continuous and Discrete Models*. Athena Scientific, Belmont, MA.

Bertsekas, D.P., Gafni, E.M. (1982). Projection methods for variational inequalities with application to the traffic assignment problem. *Mathematical Programming Study* 17, 139–159.

Bertsekas, D.P., Tsitsiklis, J.N. (1989). *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall, London.

Best, M.J., Griffin V.J. (1975). Optimisation of nonlinear performance criteria subject to flow constraints. Conference paper presented at the ORSA/TIMS National Meeting, Chicago, IL, USA, 30 April–2 May.

Bouhtou, M., van Hoesel, S., van der Kraaij, A.F., Lutton, J.-L. (2003). Tariff optimization in networks. Report RM03011, Maastricht Economic Research School on Technology and Organisation.

Boyce, D. (2004). Forecasting travel on congested urban transportation networks: Review and prospects for network equilibrium models. Conference paper presented at TRISTAN V, The Fifth Triennial Symposium on Transportation Analysis, Le Gosier, Guadeloupe, June 13–18.

Boyce, D.E., Ralevic-Dekic, B., Bar-Gera, H. (2002). Convergence of traffic assignments: How much is enough? The Delaware Valley Region Case Study. Paper presented at the 16th Annual International EMME/2 Users' Group Conference, Albuquerque, NM, March.

Brotcorne, L., Labbé, M., Marcotte, P., Savard, G. (2001). A bilevel model for toll optimization on a multicommodity transportation network. *Transportation Science* 35, 1–14.

Bruck, R.E. Jr (1975). Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. *Journal of Functional Analysis* 18, 15–26.

Bruck, R.E. Jr (1977). On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications* 61, 159–164.

Brucker, P. (1984). An O(n) algorithm for quadratic knapsack problems. *Operations Research Letters* 3, 163–166.

Busacker, R.G., Saaty, T.L. (1965). *Finite Graphs and Networks: An Introduction with Applications*. McGraw-Hill, New York.

Cascetta, E. (2001). *Transportation Systems Engineering: Theory and Methods*. *Applied Optimization*, vol. 49. Kluwer Academic, Dordrecht, The Netherlands.

Chau, C.K., Sim, K.M. (2003). The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands. *Operations Research Letters* 31, 327–334.

Chen, A., Lee, D.-H., Jayakrishnan, R. (2002). Computational study of state-of-the-art path-based traffic assignment algorithms. *Mathematics and Computers in Simulation* 2060, 1–10.

Chen, M., Bernstein, D. (2003). Solving the toll design problem with multiple user groups. *Transportation Research B* 37, 1–19.

Cohen, G. (1988). Auxiliary problem principle extended to variational inequalities. *Journal of Optimization Theory and Applications* 59, 325–333.

Cole, R., Dodis, Y., Roughgarden, T. (2003). Pricing network edges for heterogeneous selfish users. In: *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing*, 2003, pp. 521–530.

Correa, J.R., Schulz, A.S., Stier-Moses, N.E. (2004). Selfish routing in capacitated networks. *Mathematics of Operations Research* 29, 961–976.

Côté, J.-P., Marcotte, P., Savard, G. (2003). A bilevel modeling approach to pricing and fare optimization in the airline industry. *Journal of Revenue and Pricing Management* 2, 23–36.

Crouzeix, J.-P., Marcotte, P., Zhu, D. (2000). Conditions ensuring the applicability of cutting-plane methods for solving variational inequalities. *Mathematical Programming* 88, 521–539.

Dafermos, S.C. (1971). An extended traffic assignment model with applications to two-way traffic. *Transportation Science* 5, 366–389.

Dafermos, S.C. (1980). Traffic equilibrium and variational inequalities. *Transportation Science* 14, 42–54.

Dafermos, S.C. (1981). A multicriteria route-mode choice traffic equilibrium model. Unpublished manuscript, Lefschetz Center for Dynamical Systems, June.

Dafermos, S.C. (1982). The general multimodal network equilibrium problem with elastic demand. *Networks* 12, 57–72.

Dafermos, S.C., Sparrow, F.T. (1969). The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards* 73B, 91–118.

de Palma, A., Nesterov, Y. (1998). Optimization formulations and static equilibrium in congested transportation networks. CORE Discussion Paper 9861, CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Dial, R.B. (1979). A model and algorithm for multicriteria route-mode choice. *Transportation Research B* 13, 311–316.

Dial, R.B. (1996a). Bicriterion traffic assignment: Basic theory and elementary algorithms. *Transportation Science* 30, 93–111.

Dial, R.B. (1996b). Bicriterion traffic assignment: Efficient algorithms plus examples. *Transportation Research B* 31, 357–359.

Dial, R.B. (1999). Minimal-revenue congestion pricing part I: A fast algorithm for the single-origin case. *Transportation Research B* 33, 189–202.

Dial, R.B. (2000). Minimal-revenue congestion pricing part II: An efficient algorithm for the general case. *Transportation Research B* 34, 645–665.

Dontchev, A.L., Rockafellar, R.T. (2001). Ample parameterization of variational inclusions. *SIAM Journal on Optimization* 12, 170–187.

Dunn, J.C. (1973). On recursive averaging processes and Hilbert space extensions of the contraction mapping principle. *Journal of the Franklin Institute* 295, 117–133.

Escudero, L.F. (1986). A motivation for using the truncated Newton approach in a very large scale nonlinear network problem. *Mathematical Programming Study* 26, 240–244.

Evans, S.P. (1976). Derivation and analysis of some models for combining trip distribution and assignment. *Transportation Research* 10, 37–57.

Facchinei, F., Pang, J.-S. (2003a). *Finite-Dimensional Variational Inequalities and Complementarity Problems*. *Springer Series in Operations Research and Financial Engineering*, vol. I. Springer-Verlag, Berlin.

Facchinei, F., Pang, J.-S. (2003b). *Finite-Dimensional Variational Inequalities and Complementarity Problems*. *Springer Series in Operations Research and Financial Engineering*, vol. II. Springer-Verlag, Berlin.

Ferrari, P. (1995). Road pricing and network equilibrium. *Transportation Research B* 29, 357–372.

Ferris, M.C., Pang, J.-S. (1997). Engineering and economic applications of complementarity problems. *SIAM Review* 39, 669–713.

Fisk, C.S. (1980). Some developments in equilibrium traffic assignment. *Transportation Research B* 14, 243–255.

Florian, M. (1977). A traffic equilibrium model of travel by car and public transit modes. *Transportation Science* 11, 166–179.

Florian, M. (1979). Asymmetrical variable demand multi-mode traffic equilibrium problems: Existence and uniqueness of solutions and a solution algorithm. Publication 347, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, Canada.

Florian, M., Hearn, D.W. (1995). Network equilibrium models and algorithms. In: Ball, M.O., Magnanti, T.L., Monma, C.L., Nemhauser, G.L. (Eds.), *Network Routing*. *Handbooks in Operations Research and Management Science*, vol. 8. North-Holland, Amsterdam, pp. 485–550, Chapter 6.

Florian, M., Wu, J.H., He, S. (2002). Multi-mode variable demand network equilibrium model with hierarchical logit structure. In: Gendreau, M., Marcotte, P. (Eds.), *Transportation and Network Analysis: Current Trends*. Kluwer Academic, Dordrecht, The Netherlands, pp. 119–133.

Frank, M., Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly* 3, 95–110.

Friesz, T.L., Cho, H.-J., Metha, N.J., Tobin, R.L., Anandalingam, G. (1992). A simulated annealing approach to the network design problem with variational inequality constraints. *Transportation Science* 26, 18–26.

Gabriel, S.A., Bernstein, D. (1997). The traffic equilibrium problem with non-additive path costs. *Transportation Science* 31, 337–348.

Gafni, E.M., Bertsekas, D.P. (1984). Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization* 22, 936–964.

Gallager, R.G. (1977). Loops in multicommodity flows. In: *Proceedings of the 10th IEEE Conference on Decision and Control*, New Orleans, TX, pp. 819–825.

Gartner, N.H. (1980). Optimal traffic assignment with elastic demands: A review. Part II: Algorithmic approaches. *Transportation Science* 14, 192–208.

Glowinski, R., Lions, J.-L., Trémolières, R. (1981). *Numerical Analysis of Variational Inequalities*. North-Holland, Amsterdam.

Goffin, J.-L., Marcotte, P., Zhu, D.L. (1997). An analytic center cutting-plane method for pseudomonotone variational inequalities. *Operations Research Letters* 20, 1–6.

Hagstrom, J.N., Tseng, P. (1998). Traffic equilibrium: Link flows, path flows and weakly/strongly acyclic solutions. Technical Report M/C 294, Department of Information and Decision Sciences, University of Illinois, Chicago, IL. Available at http://www.math.washington.edu/~tseng/papers.html.

Hamdouch, Y., Marcotte, P., Nguyen, S. (2004a). Capacitated transit assignment with loading priorities. *Mathematical Programming B* 101, 205–230.

Hamdouch, Y., Marcotte, P., Nguyen, S. (2004b). A strategic model for dynamic traffic assignment. *Networks and Spatial Economics* 4, 291–315.

Hammond, J.H. (1984). Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms. PhD thesis, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA.

Hartman, P., Stampacchia, G. (1966). On some non-linear elliptic differential-functional equations. *Acta Mathematica* 115, 271–310.

Hearn, D.W. (1980). Bounding flows in traffic assignment models. Research Report 80-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL.

Hearn, D.W., Ramana, M.V. (1998). Solving congestion toll pricing models. In: Marcotte, P., Nguyen, S. (Eds.), *Equilibrium and Advanced Transportation Modelling*. Kluwer Academic, Boston, MA, pp. 109–124.

Hearn, D.W., Lawphongpanich, S., Ventura, J.A. (1987). Restricted simplicial decomposition: Computation and extensions. *Mathematical Programming Study* 31, 99–118.

Heydecker, B.G. (1986). On the definition of traffic equilibrium. *Transportation Research B* 20, 435–440.

Huang, H.-J., Bell, M.G.H. (1998). Continuous equilibrium network design problem with elastic demand: Derivative-free solution methods. In: Bell, M.G.H. (Ed.), *Transportation Networks: Recent Methodological Advances*. Pergamon Press, Amsterdam, pp. 175–183.

Iusem, A.N. (1998). On some properties of paramonotone operators. *Journal of Convex Analysis* 5, 269–278.

Jayakrishnan, R., Tsai, W.K., Prashker, J.N., Rajadhyaksha, S. (1994). Faster path-based algorithm for traffic assignment. *Transportation Research Record* 1443, 75–83.

Jorgensen, N.O. (1963). Some aspects of the urban traffic assignment problem. Master's thesis, Institute of Transportation and Traffic Engineering, University of California, Berkeley, CA.

Josefsson, M., Patriksson, M. (2005). On the applicability of sensitivity analysis formulas for traffic equilibrium models. In: Hearn, D., Lawphongpanich, S., Smith, M. (Eds.), *Mathematical and Computational Methods for Congestion Charging*. Proceedings of the Theory and Practice of Congestion Charging Symposium, Imperial College, London, August 18–20, 2003. *The Applied Optimization Series*, vol. 148. Springer-Verlag, Berlin, pp. 117–141.

Julsain, H. (1998). Tarification dans les réseaux de télécommunications: une approche par programmation mathématique à deux niveaux. Master's thesis, École Polytechnique, Montréal.

Kennington, J.L., Helgason, R.V. (1980). *Algorithms for Network Programming*. Wiley, New York.

Klincewicz, J.G. (1983). A Newton method for convex separable network flow problems. *Networks* 13, 427–442.

Knight, F.H. (1924). Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38, 582–606.

Konnov, I.V. (2001). *Combined Relaxation Methods for Variational Inequalities*. *Lecture Notes in Economics and Mathematical Systems*, vol. 495. Springer-Verlag, Berlin.

Korpelevich, G.M. (1977). The extragradient method for finding saddle points and other problems. *Matecon* 13, 35–49.

Kyparisis, J. (1990). Solution differentiability for variational inequalities. *Mathematical Programming* 48, 285–301.

Labbé, M., Marcotte, P., Savard, G. (1998). A bilevel model of taxation and its application to optimal highway pricing. *Management Science* 44, 1595–1607.

Labbé, M., Marcotte, P., Savard, G. (1999). On a class of bilevel programs. In: Di Pillo, G., Giannessi, F. (Eds.), *Nonlinear Optimization and Related Topics*. Kluwer Academic, Dordrecht, The Netherlands, pp. 183–206.

Larsson, T., Patriksson, M. (1992). Simplicial decomposition with disaggregated representation for the traffic assignment problem. *Transportation Science* 26, 4–17.

Larsson, T., Patriksson, M. (1994a). A class of gap functions for variational inequalities. *Mathematical Programming* 64, 53–79.

Larsson, T., Patriksson, M. (1994b). Equilibrium characterizations of solutions to side constrained asymmetric traffic assignment models. *Le Matematiche* 49, 249–280.

Larsson, T., Patriksson, M. (1995). An Augmented Lagrangean dual algorithm for link capacity side constrained traffic assignment problems. *Transportation Research B* 29, 433–455.

Larsson, T., Patriksson, M. (1997). Price-directive traffic management – an approach utilizing side constrained traffic equilibrium models. *Rendiconti del Circolo Matematico di Palermo, Serie II* 48, 147–170.

Larsson, T., Patriksson, M. (1998). Side constrained traffic equilibrium models – traffic management through link tolls. In: Marcotte, P., Nguyen, S. (Eds.), *Equilibrium and Advanced Transportation Modelling*. Kluwer Academic, New York, pp. 125–151.

Larsson, T., Patriksson, M. (1999). Side constrained traffic equilibrium models: Analysis, computation and applications. *Transportation Research B* 33, 233–264.

Larsson, T., Liu, Z.-W., Patriksson, M. (1997). A dual scheme for traffic assignment problems. *Optimization* 42, 323–358.

Larsson, T., Patriksson, M., Strömberg, A.-B. (1999). Ergodic, primal convergence in dual subgradient schemes for convex programming. *Mathematical Programming* 86, 283–312.

Larsson, T., Lundgren, J., Patriksson, M., Rydergren, C. (2001). Most likely traffic equilibrium route flows: Analysis and computation. In: Giannessi, F., Maugeri, A., Pardalos, P.M. (Eds.), *Equilibrium Problems: Nonsmooth Optimization and Variational Inequality Models*. Proceedings of an International Workshop in Memory of Marino De Luca, Hotel Villa Diodoro, Taormina, Italy, December 3–5, 1998. Kluwer Academic, Dordrecht, The Netherlands, pp. 129–159.

Larsson, T., Lindberg, P.-O., Lundgren, J., Patriksson, M., Rydergren, C. (2002). On traffic equilibrium models with a nonlinear time/money relation. In: Patriksson, M., Labbé, M. (Eds.), *Transportation Planning: State of the Art*. Kluwer Academic, Dordrecht, The Netherlands, pp. 19–31.

Larsson, T., Patriksson, M., Rydergren, C. (2004). A column generation procedure for the side constrained traffic equilibrium problem. *Transportation Research B* 38, 17–38.

Lawphongpanich, S., Hearn, D.W. (1984). Simplicial decomposition of the asymmetric traffic assignment problem. *Transportation Research B* 18, 123–133.

LeBlanc, L.J., Morlok, E.K., Pierskalla, W.P. (1975). An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research* 9, 308–318.

Leurent, F. (1993). Cost versus time equilibrium over a network. *European Journal of Operational Research* 71, 205–221.

Leventhal, T., Nemhauser, G., Trotter, L. Jr. (1973). A column generation algorithm for optimal traffic assignment. *Transportation Science* 7, 168–176.

Lim, A.C. (2002). Transportation network design problems: An MPEC approach. PhD thesis, Department of Mathematical Sciences, The Johns Hopkins University Baltimore, MD.

Luo, Z.-Q., Pang, J.-S., Ralph, D. (1996). *Mathematical Programs with Equilibrium Constraints*. Cambridge Univ. Press, Cambridge, UK.

Marcotte, P. (1986a). A new algorithm for solving variational inequalities, with application to the traffic assignment problem. *Mathematical Programming* 33, 339–351.

Marcotte, P. (1986b). Network design problem with congestion effects: A case of bilevel programming. *Mathematical Programming* 34, 142–162.

Marcotte, P. (1991). Application of Khobotov's algorithm to variational inequalities and network equilibrium problems. *INFOR* 29, 258–270.

Marcotte, P. (1998). Reformulations of a bicriterion equilibrium problem. In: Fukushima, M., Qi, L. (Eds.), *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*. Kluwer Academic, Dordrecht, The Netherlands, pp. 269–292.

Marcotte, P., Guélat, J. (1988). Adaptation of a modified Newton method for solving the asymmetric traffic equilibrium problem. *Transportation Science* 22, 112–124.

Marcotte, P., Marquis, G. (1992). Efficient implementation of heuristics for the continuous network design problem. *Annals of Operations Research* 34, 163–176.

Marcotte, P., Nguyen, S. (1998). Hyperpath formulations of traffic assignment problems. In: Marcotte, P., Nguyen, S. (Eds.), *Equilibrium and Advanced Transportation Modelling*. Kluwer Academic, Dordrecht, The Netherlands, pp. 175–200.

Marcotte, P., Savard, G. (2001). Bilevel programming: Formulation, applications, algorithms. In: Floudas, C.A., Pardalos, P.M. (Eds.), *The Encyclopedia of Optimization*. Kluwer Academic, Dordrecht, The Netherlands, pp. 155–159.

Marcotte, P., Savard, G. (2002). A bilevel programming approach to price setting. In: Zaccour, G. (Ed.), *Decision and Control in Management Science. Essays in Honor of Alain Haurie*. Kluwer Academic, Dordrecht, pp. 97–117.

Marcotte, P., Wynter, L. (2004). A new look at the multi-class network equilibrium problem. *Transportation Science* 38, 282–292.

Marcotte, P., Zhu, D.L. (1997). Equilibria with infinitely many differentiated classes of customers. In: Pang, J.-S., Ferris, M. (Eds.), *Complementarity and Variational Problems: State of the Art*. SIAM, Philadelphia, pp. 234–258.

Marcotte, P., Nguyen, S., Tanguay, K. (1996). Implementation of an efficient algorithm for the multiclass traffic assignment problem. In: Lesort, J.-B. (Ed.), *Transportation and Traffic Theory*. Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France, 24–26 July, 1996. Elsevier, Tarrytown, NY, pp. 217–236.

Marcotte, P., Nguyen, S., Schoeb, A. (2004). A strategic model of traffic assignment in static capacitated networks. *Operations Research* 52, 191–212.

Miller, S.D., Payne, H.J., Thompson, W.A. (1975). An algorithm for traffic assignment on capacity constrained transportation networks with queues. Paper presented at the Johns Hopkins Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore, MD, April 2–4.

Murchland, J.D. (1970). Road network traffic distribution in equilibrium. In: Henn, R., Künzi, H.P., Schubert, H. (Eds.), *Mathematical Models in the Social Sciences, vol. 8*. II Oberwolfach-Tagungüber Operations Research, Mathematisches Forschungsinstitut, Oberwolfach, 20–25 October, 1969. Anton Hain Verlag, Meisenheim am Glan, pp. 145–183 (in German). Translation by H.A. Paul.

Nagurney, A., Dong, J. (2002). *Supernetworks: Decision-Making for the Information Age*. Edward Elgar, Cheltenham, UK.

Nesterov, Y. (2000). Stable traffic equilibria, properties and applications. *Optimization and Engineering* 3, 29–50.

Nguyen, S. (1974). An algorithm for the traffic assignment problem. *Transportation Science* 8, 203–216.

Nguyen, S., Dupuis, C. (1984). An efficient method for computing traffic equilibria in networks with asymmetric transportation costs. *Transportation Science* 18, 185–202.

Nguyen, S., Pallottino, S. (1988). Equilibrium traffic assignment for large scale transit networks. *European Journal of Operational Research* 37, 176–186.

Nguyen, S., Pallottino, S., Malucelli, F. (2001). A modeling framework for the passenger assignment on a transport network with time-tables. *Transportation Science* 35, 238–249.

Nguyen, S., Pallottino, S., Scutellá, M.G. (2002). A new dual algorithm for shortest path reoptimization. In: Gendreau, M., Marcotte, P. (Eds.), *Transportation and Network Analysis: Current Trends*. Kluwer Academic, New York, pp. 221–235.

Pallottino, S., Scutellà, M.G. (2003). A new algorithm for reoptimizing shortest paths when the arc costs change. *Operations Research Letters* 31, 149–160.

Patriksson, M. (1994a). On the convergence of descent methods for monotone variational inequalities. *Operations Research Letters* 16, 265–269.

Patriksson, M. (1994b). *The Traffic Assignment Problem: Models and Methods*. *Topics in Transportation*. VSP BV, Utrecht, The Netherlands.

Patriksson, M. (1998). *Nonlinear Programming and Variational Inequalities: A Unified Approach*. Kluwer Academic, Dordrecht, The Netherlands.

Patriksson, M. (2004). Sensitivity analysis of traffic equilibria. *Transportation Science* 38, 258–281.

Patriksson, M. (2006). *Traffic Equilibrium Models: Analysis, Applications, and Optimization Algorithms*. Springer-Verlag, in preparation.

Patriksson, M., Josefsson, M. (2003). Sensitivity analysis of separable traffic equilibrium solutions, with application to bilevel optimization in network design. Report, Department of Mathematics, Chalmers University of Technology, Gothenburg, Sweden. *Transportation Research B*, in press.

Patriksson, M., Rockafellar, R.T. (2002). A mathematical model and descent algorithm for bilevel traffic management. *Transportation Science* 36, 271–291.

Patriksson, M., Rockafellar, R.T. (2003). Sensitivity analysis of variational inequalities over aggregated polyhedra, with application to traffic equilibria. *Transportation Science* 37, 56–68.

Payne, H.J., Thompson, W.A. (1975). Traffic assignment on transportation networks with capacity constraints and queueing. Paper presented at the 47th National ORSA Meeting/TIMS 1975 North-American Meeting, Chicago, IL, April 30–May 2.

Robinson, S.M. (1980). Strongly regular generalized equations. *Mathematics of Operations Research* 5, 43–62.

Robinson, S.M. (1985). Implicit B-differentiability in generalized equations. Technical Summary Report 2854, Mathematics Research Center, University of Wisconsin at Madison, Madison, WI.

Robinson, S.M. (1991). An implicit-function theorem for a class of nonsmooth functions. *Mathematics of Operations Research* 16, 292–309.

Rockafellar, R.T. (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* 14, 877–898.

Rockafellar, R.T. (1984). *Network Flows and Monotropic Optimization*. Wiley, New York. Also published by Athena Scientific, Belmont, MA, 1998.

Roughgarden, T., Tardos, É. (2002). How bad is selfish routing? *Journal of the ACM* 49, 236–259.
Sheffi, Y. (1985). *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice Hall, Englewood Cliffs, NJ.
Shimizu, K., Ishizuka, Y., Bard, J.F. (1997). *Nondifferentiable and Two-Level Programming*. Kluwer Academic, Boston, MA.
Smith, M.J. (1979). The existence, uniqueness and stability of traffic equilibria. *Transportation Research B* 13, 295–304.
Smith, M.J. (1984). Two alternative definitions of traffic equilibrium. *Transportation Research B* 18, 63–65.
Solodov, M.V. (2003). Convergence rate analysis of iterative algorithms for solving variational inequality problems. *Mathematical Programming* 96, 513–528.
Solodov, M.V., Svaiter, B.F. (1999a). A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis* 7, 323–345.
Solodov, M.V., Svaiter, B.F. (1999b). A hybrid projection-proximal point algorithm. *Journal of Convex Analysis* 6, 59–70.
Spiess, H., Florian, M. (1989). Optimal strategies: A new assignment model for transit networks. *Transportation Research B* 23, 83–102.
Suwansirikul, C., Friesz, T.L., Tobin, R.L. (1987). Equilibrium decomposed optimization: A heuristic for the continuous equilibrium network design problem. *Transportation Science* 21, 254–260.
Tobin, R.L., Friesz, T.L. (1988). Sensitivity analysis for equilibrium network flow. *Transportation Science* 22, 242–250.
Toint, Ph., Wynter, L. (1996). Asymmetric multiclass traffic assignment: A coherent formulation. In: Lesort, J.-B. (Ed.), *Transportation and Traffic Theory*. Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France, 24–26 July, 1996. Pergamon Press, Oxford, pp. 237–260.
Wang, Y.J., Xiu, N.H., Wang, C.Y. (2001). Unified framework of extragradient-type methods for pseudomonotone variational inequalities. *Journal of Optimization Theory and Applications* 111, 641–656.
Watling, D. (1999). Stability of the stochastic equilibrium assignment problem: a dynamical systems approach. *Transportation Research B* 33, 281–312.
Yang, H., Huang, H.-J. (2004). The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. *Transportation Research B* 38, 1–15.
Zhu, D.L., Marcotte, P. (1996). Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization* 6, 714–726.
Zuhovickiĭ, S.I., Polyak, R.A., Primak, M.E. (1969). Two methods of search for equilibrium points of *n*-person concave games. *Soviet Mathematics Doklady* 10, 279–282.

C. Barnhart and G. Laporte (Eds.), *Handbook in OR & MS, Vol. 14*

DOI: 10.1016/S0927-0507(06)14011-6

Chapter 11

ITS and Traffic Management

M. Papageorgiou

Dynamic Systems and Simulation Laboratory, Technical University of Crete, 731 00, Chania, Crete, Greece
E-mail: markos@dssl.tuc.gr

M. Ben-Akiva

Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA
E-mail: mba@mit.edu

J. Bottom

Charles River Associates, Inc., Boston, MA, USA
E-mail: jbottom@crai.com

P.H.L. Bovy

Transportation and Traffic Planning Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands
E-mail: p.h.l.bovy@tudelft.nl

S.P. Hoogendoorn

Transportation and Traffic Planning Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands
E-mail: s.p.hoogendoorn@tudelft.nl

N.B. Hounsell

Transportation Research Group, School of Civil Engineering and the Environment, University of Southampton, Hants, SO17 1BJ, UK
E-mail: N.B.Hounsell@soton.ac.uk

A. Kotsialos

School of Engineering, University of Durham, South Road, Durham, DH1 3LE, UK
E-mail: apostolos.kotsialos@durham.ac.uk

M. McDonald

Transportation Research Group, School of Civil Engineering and the Environment, University of Southampton, Hants, SO17 1BJ, UK
E-mail: M.Mcdonald@soton.ac.uk

1 Introduction

The observed traffic conditions on road and highway networks result from a quite complex-to-describe confrontation of supply and demand. Supply is

mainly determined from the available road and highway infrastructure, most notably its capacity. Demand is the collective outcome of individual driver decisions regarding the effectuation (or not) of a trip, the choice of transportation mode, of departure time, of the route to be followed, etc. Traffic congestion is observed in increasing levels on road and highway networks around the world, with detrimental consequences for traffic efficiency, safety as well as for the environment. Traffic congestion affects the nominal capacity of the available infrastructure leading to a vicious cycle of further congestion increase, further infrastructure degradation, and so forth. In fact, the traffic throughput measured in congested road or highway networks is usually well below the nominal network capacity. Traffic control measures and strategies described in this chapter aim at maintaining the available infrastructure capacity close to nominal levels, protecting the traffic networks from the detrimental effects of oversaturation and even gridlock. In this sense, traffic control is deemed to mainly act on the supply side of the basic traffic equation. Other operational measures have been employed in an attempt to reduce congestion by influencing the manifest traffic demand; this includes various forms of administrative restrictions or of demand management (road pricing being the most prominent), which, however, are not addressed in this chapter.

The chapter consists of 4 main overview sections, Section 2 presenting an overview of traffic flow modeling advancements, Section 3 addressing the issue of route guidance and information systems, while Sections 4 and 5 are concerned with specific road and motorway network control systems, respectively.

2 Traffic flow modeling

2.1 Traffic flow modeling approaches

Understanding traffic flow characteristics (e.g., headway distributions, relation between density and speed, capacity distributions) and knowledge of the associated analytical tools (e.g., queuing models, shockwave theory, simulation models) to predict the dynamics of these characteristics under given demand, supply and control conditions, is an essential requirement for the planning, design and operation of a transportation system. For the analysis of a simple arterial, on-ramp, or merge area, as well as for studying traffic flow operations in urban or motorway networks, being able to predict the traffic performance is an essential factor in the analysis of the system.

Traffic flow modeling research started when Lighthill and Whitham (1955) presented their seminal paper on the wave dynamics of traffic flow. Their work was based on the analogy of vehicles in traffic flow and particles in a fluid. Since then, the mathematical description of traffic flow has been a lively subject of research and debate for traffic engineers. This has resulted in a broad scope of low-end and high-end models.

Before giving an overview of these models, we need to emphasize that it is highly unlikely that traffic science will ever produce a complete theory on the motion of individual cars. Despite of this, the last five decades have provided tools to construct a framework of useful – albeit incomplete – theories from traffic observations and experimentation. These incomplete theories are neither deductive (i.e., stemming from excellent theories), nor inductive (black box), but rather intermediate; basic mathematical model structures are adopted, after which specific flow properties are determined from empirical or experimental data. Since the only accurate physical law in traffic flow theory is the conservation of vehicle's equation, the main challenge of traffic flow researchers is to look for intuitive and useful theories of traffic flow.

2.1.1 Microscopic and macroscopic characteristics and models

A key distinction made in the study of traffic systems is that between *microscopic* and *macroscopic* variables. Microscopic characteristics (e.g., time headways, individual speeds and distance headways) pertain to the individual driver–vehicle unit in relation to the other drivers in the flow. Microscopic models describe the behavior of individual vehicles in relation to the infrastructure and other vehicles in the flow. On the contrary, macroscopic characteristics pertain to the properties of the traffic flow as a whole (for instance at a cross-section, or at a time instant). Examples of macroscopic characteristics are flow, time-mean speed, density, and space-mean speed. Macroscopic models describe traffic flow in terms of its macroscopic characteristics. The intermediate *mesoscopic* level is used to indicate the behavior of groups of drivers.

2.1.2 Properties of traffic flow

Microscopic and macroscopic characteristics of traffic flow have been studied for many years. Studies concern a large variety of aspects such as the distribution of headways, statistical relations between speed and density, capacity of the infrastructure, propagation of shock-waves, etc. A lot of consideration has been put into average behavior of drivers under the assumption of stationary flow conditions. Under these conditions, it is reasonable to assume that the average behavior of drivers is the same for the same average conditions. That is, drivers having a certain speed u, will on average maintain the same distance headway $s = 1/k$ (where k denotes the vehicular density, i.e., the mean number of vehicles per unit roadway length) with respect to the preceding vehicles. This in turn implies that if we may assume that there exists some *statistical* (*but not necessarily causal!*) relation among the density k (or equivalently, the mean distance headway s), the (space) mean speed u and flow $q = ku$, then it holds:

$$q = Q(k) = kU(k). \tag{1}$$

Figure 1 shows an example of the three forms of this fundamental relation, showing some of its important properties ($\mathrm{d}U/\mathrm{d}k < 0$, $U(k_j) = 0$, $Q(k_j) = 0$, $q_c = \max_k Q(k)$, etc.). The *fundamental relation* will depend on

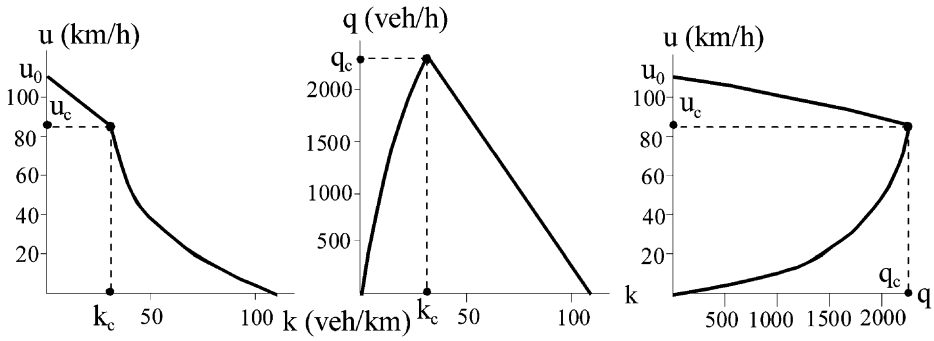


Fig. 1. Examples of the fundamental relations between flow, density, and speed.

the different properties of the road (width of the lanes, grade), flow composition (percentage of trucks, fraction of commuters, experienced drivers, etc.), external conditions (weather and ambient conditions), traffic regulations, etc.

Traffic flow observations however show that many data are not on the fundamental diagram. While some of these points can be explained by stochastic fluctuations (e.g., vehicles have different sizes, drivers have different desired speeds and following distances), a number of researchers have suggested that these differences are structural and stem from the dynamic properties of traffic flow. That is, they reflect so-called transient states (i.e., changes from congestion to free flow (acceleration phase) or from free flow to congestion (deceleration phase)) of traffic flow.

Several authors have studied the nonlinear or even chaotic-like behavior of the traffic system (cf. Bovy and Hoogendoorn, 2000; Pozybill, 1998). Among these behaviors are hysteresis and metastable or unstable behavior of traffic flow. The latter implies that in heavy traffic a critical disturbance may be amplified and develop into a traffic jam (spontaneous phase-transitions). In illustration, empirical experiments performed in Forbes et al. (1958), and Edie and Foote (1958, 1960) have shown that a disturbance at the foot of an upgrade propagates from one vehicle to the next, while being amplified until at some point a vehicle came to a complete stop. This *instability effect* implies that once the density crosses some critical value, traffic flow becomes rapidly more congested *without any obvious reasons*. More empirical evidence of this instability and start–stop wave formation can be found in among others (Verweij, 1985; Ferrari, 1989; Leutzbach, 1991). In Kerner and Rehborn (1997) and Kerner (1999) it is empirically shown that local jams can persist for several hours, while maintaining their form and characteristic properties. In other words, the stable complex structure of a traffic jam can and does exist on motorways.[1] These findings show that traffic flow has some *chaotic-like properties*, implying that

[1] Apart from the formation of stop-and-go waves and localized structures, a hysteric phase-transition from free-traffic to *synchronized flow* that mostly appears near on-ramps is described in Kerner and

microscopic disturbances in the flow can result in the on-set of local traffic jams persisting for several hours.

Having said this, it should be clear that traffic flow shows some interesting phenomena, which must be reflected correctly by the different models that have been proposed. The remainder of this section focuses on these different models, while discussing their most important properties.

2.1.3 Approaches to traffic flow modeling

Traffic flow models may be categorized using various dimensions (deterministic or stochastic, continuous or discrete, analytical or simulation, etc.), stochastic. The most common classification is the distinction between microscopic and macroscopic traffic flow modeling approaches. However, this distinction is not unambiguous, due to the existence of hybrid models. This is why below, models are categorized based on the following aspects:

1. *Representation* of the traffic flow in terms of flows (macroscopic), groups of drivers (mesoscopic), or individual drivers (microscopic).
2. *Underlying behavioral theory*, which can be based on characteristics of the flow (macroscopic), or individual drivers (microscopic behavior).

The remainder of this section uses this classification to discuss some important flow models. Table 1 presents an overview of these models and the relevant sections.

2.1.4 Microscopic traffic flow models

A microscopic model provides a description of the movements of individual vehicles that are considered to be a result of the characteristics of drivers and vehicles, the interactions between driver–vehicle units, the interactions between driver–vehicle units and the road characteristics, external conditions, and the traffic regulations and control. In general, two types of driver tasks are distinguished: longitudinal tasks (acceleration, maintaining speed, distance keeping relative to leading vehicle) and lateral tasks (lane changing, overtaking). With respect to the longitudinal movement, most microscopic simulation

Table 1.
Overview of traffic flow model classification

Representation	Behavioral rules	
	Microscopic	Macroscopic
Vehicle-based	Microscopic flow models (2.1.4)	Particle models (2.1.5)
Flow-based	Gas-kinetic models (2.1.7)	Macroscopic models (2.1.6)

Rehborn (1997). In addition, transitions from synchronized flow to the jammed traffic state occur in congestion, upstream of the bottleneck.

models assume that a driver will only respond to the one vehicle (the leader) that is driving in the same lane, directly in front of her.

When the number of driver–vehicle units on the road is very small, the driver can freely choose her speed given her preferences and abilities, the roadway conditions, curvature, prevailing speed-limits, etc. In any case, there will be little reason for the driver to adapt her speed to the other road-users. The target-speed of the driver is the so-called free speed. In real life, the free speed will vary from one driver to another, but also the free speed of a single driver may change over time. Most microscopic models assume however that the free speeds have a constant value that is driver-specific. When traffic becomes denser, drivers will no longer be able to choose the speed freely, since they will not always be able to overtake a slower vehicle. The driver will need to adapt her speed to the prevailing traffic conditions, i.e., the driver is following. In the remainder, we will discuss some of these car-following models. Models for the lateral tasks, such as deciding to perform a lane-change and gap-acceptance, will not be discussed in this section in detail; a concise framework of lane changing modeling is provided by Ahmed et al. (1996).

Safe-distance models. The first car-following models were developed in Pipes (1953) and were based on the assumption that drivers maintain a safe distance: a good rule for following vehicle $i-1$ at a safe distance s_i is to allow at least the length S_0 of a car between vehicle i and the vehicle ahead for every ten miles per hour of speed v_i at which i is traveling:

$$s_i = S(v_i) = S_0 + T_\mathrm{r} \cdot v_i, \tag{2}$$

where S_0 is the effective length of a stopped vehicle (including additional distance in front), and T_r denotes a parameter (comparable to the reaction time). A similar approach was proposed in Forbes et al. (1958). Both theories were compared to field measurements. It was concluded that according to Pipes' theory, the minimum headways are slightly less at low and high velocities than observed in empirical data. However, considering the models' simplicity, agreement with real-life observations was amazing (cf. Pignataro, 1973).

In Leutzbach (1988) a more refined model describing the spacing of constrained vehicles in the traffic flow was proposed. Considering an overall reaction time T_r, the distance needed to come to a full stop given the initial speed v_i, the friction coefficient μ, and gravity g, equals

$$S(v_i) = S_0 + T_\mathrm{r} \cdot v_i + \frac{v_i^2}{2\mu g}. \tag{3}$$

Stimulus-response models. Stimulus-response models are dynamic models that describe the reaction of drivers as a function of changes in distance, speeds, etc., relative to the vehicle in front. These models are applicable to relatively busy traffic flows, where the overtaking possibilities are small and drivers are obliged to follow the vehicle in front of them. Drivers do not want

the gap in front of them to become too large, so that other drivers might enter it. At the same time, the drivers will generally be inclined to keep a safe distance.

Stimulus-response models assume that drivers control their acceleration. The well-known model proposed in Chandler et al. (1958) is based on the intuitive hypothesis that a driver's acceleration is proportional to the relative speed $v_{i-1} - v_i$:

$$a_i(t + T_r) = \dot{v}_i(t) = \alpha\big(v_{i-1}(t) - v_i(t)\big), \tag{4}$$

where T_r again denotes the overall reaction time, and α denotes the sensitivity. Based on field experiments, conducted to quantify the parameter values for the reaction time T_r and the sensitivity α, it was concluded that α depended on the distance between the vehicles: when the vehicles were close together, the sensitivity was high, and vice versa. The following specification was proposed by

$$\alpha = \frac{\alpha_0}{x_{i-1}(t) - x_i(t)}. \tag{5}$$

One of the main aspects of a dynamic model is its stability, i.e., whether small disturbances will damp out *or* be amplified. For the stimulus-response model (4), two types of stability can be distinguished, namely *local stability* (stability of response of a driver on the leading vehicle $i - 1$), and *asymptotic stability* (propagation of disturbances along a platoon). Asymptotic stability is of more practical importance than local stability. If a platoon of vehicles is asymptotically unstable, a small disturbance in the movement of the first vehicle is amplified as it is passed over to the next vehicle, which in turn can lead to dangerous situations. Let us briefly consider both kinds of stability. The local and asymptotic stability of the model depends on the sensitivity α and the reaction time T_r, i.e., the model is *locally stable* if $C = \alpha T_r < \pi/2$. *Asymptotic stability* requires $C = \alpha T_r < 1/2$. Note that local stability is less critical than asymptotic stability because the stimulus-response model becomes unstable only for (unrealistically) large response times or large sensitivity values.

This simple model has several undesirable and unrealistic properties. For one, vehicles tend to get *dragged along* when the vehicle in front is moving at a higher speed. Furthermore, when the distance $s_i(t)$ is very large, the speeds can become unrealistically high. To remedy this deficiency, sensitivity α can be defined as a decreasing function of the distance. In more general terms, the sensitivity thus can be defined as follows

$$\alpha = \frac{\alpha_0 (v_i(t + T_r))^m}{(x_{i-1}(t) - x_i(t))^l}. \tag{6}$$

Equation (6) implies that the following vehicle adjusts its speed $v_i(t)$ proportionally to both distances and speed differences with delay T_r. The extent to which this occurs depends on the values of α, l, and m. Combining Equations (4) and (6), and integrating the result, relations between the speed $v_i(t + T_r)$

and the distance headway $x_{i-1}(t)-x_i(t)$ can be determined. Assuming stationary traffic conditions, the following relation between the *equilibrium speed* U and the density k results

$$U(k) = U^0\left(1 - \left(\frac{k}{k_j}\right)^{(l-1)}\right)^{1/(1-m)} \tag{7}$$

for $m \neq 1$ and $l \neq 1$; $k = 1/(x_{i-1} - x_i)$ denotes the density (average number of vehicles per unit roadway length); k_j is the so-called *jam-density* (density at which $U = 0$); U^0 is the mean free speed (at $k = 0$). We refer to Leutzbach (1988) for a more general expression of (7).

An alternative approach was proposed in Helly (1959), which includes an additional term describing the tendency of drivers to maintain a certain desired following distance $S_i(t)$:

$$a_i(t + T_\mathrm{r}) = \alpha_1\big(v_{i-1}(t) - v_i(t)\big) + \alpha_2\big(x_{i-1}(t) - x_i(t) - S_i(t)\big), \tag{8}$$

where

$$S_i(t) = \beta_0 + \beta_1 v_i(t) + \beta_2 a_i(t), \quad \text{where } \beta_j \geqslant 0. \tag{9}$$

Car-following models have been mainly applied to single lane traffic (e.g., tunnels, cf. Newell (1961)) and traffic stability analysis (Herman et al., 1959; May, 1990). The parameters of the model (7) have been estimated using macroscopic and microscopic data by a large number of researchers. It should be noted that no generally applicable set of parameter estimates has been found so far, i.e., estimates are site-specific. An overview of parameter estimates can be found in Brackstone and McDonald (1999).

Optimal speed models. So far, the models considered mainly describe the car-following task where the follower (in time) will aim to drive at the speed of the leader, at a certain distance gap. Of course, there can be choices of the desired speed other than the speed of the leader. In Bando et al. (1995) it is assumed that the desired speed is a function of the distance between the vehicles under consideration, i.e.,

$$a_i(t) = \frac{U_\mathrm{des}(x_{i-1}(t) - x_i(t)) - v_i(t)}{T_\mathrm{r}}, \tag{10}$$

where $U_\mathrm{des}(x_{i-1} - x_i) = U_0 \tanh(x_{i-1} - x_i)$.

Psycho-spacing models. The car-following models discussed so far have a mechanistic character. The only human element is the presence of a finite reaction time T_r. However, in reality a driver is not able to:

- observe a stimulus lower than a given value (perception threshold);
- evaluate a situation and determine the required response precisely, for instance due to observation errors resulting from radial motion observation;

- manipulate the gas and brake pedal precisely.

Furthermore, due to the need to distribute her attention to different tasks, a driver will generally not be permanently occupied with the car-following task. This type of considerations has inspired a different class of car-following models, namely the *psycho-spacing models*. The first psycho-spacing models were based on theories borrowed from perceptual psychology provided in Michaels (1963); cf. Leutzbach and Wiedemann (1986). In these models, car-following behavior is described using a plane with relative speed and headway distance as axes. The model is illustrated in Figure 2.

It is assumed that the vehicle in front has a constant speed and that the potential car-following driver catches up with a constant relative speed $v'_{\mathrm{r}} = v_{i-1} - v_i$. As long as the headway distance $x_{i-1} - x_i$ is larger than s_g, there is no response. If the absolute value of the relative speed is smaller than a boundary value v_{rg}, then there is also no response because the driver cannot perceive the relative speed. The threshold value is not a constant but depends on the speed difference. If the vehicle crosses the boundary, it responds with a constant positive or negative acceleration. This happens in Figure 2 first at point A, then at point C, then point B, etc. The term *pendeling* (the pendulum of a clock) for the fact that the distance headway varies around a constant value, even if the vehicle in front has a constant speed, has been introduced in Leutzbach (1988). In this action-point model the size of the acceleration is arbitrary in the first instance, whereas it was the main point of the earlier discussed car-following models.

Action point models form the basis for a large number of contemporary microscopic traffic flow models. In Brackstone and McDonald (1999) it is concluded that it is hard to come to a definitive conclusion on the validity of these models, mainly because the calibration of its elements has not been successful so far.

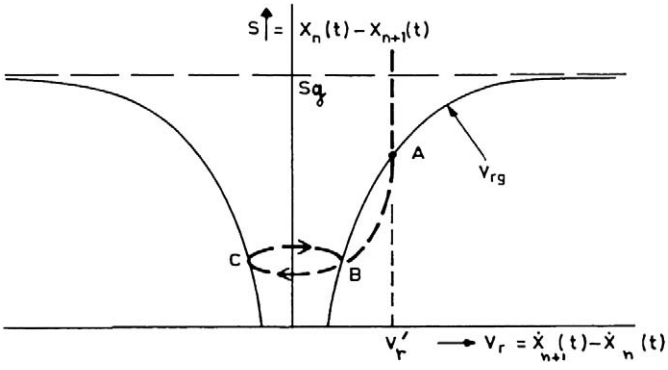


Fig. 2. Basic action-point car-following model.

Submicroscopic simulation models. In addition to describing the time–space behavior of the individual entities in the traffic system, submicroscopic simulation models describe the functioning of specific parts and processes of vehicles and driving tasks. For instance, a submicroscopic simulation model describes the way in which a driver applies the brakes, considering among other things the driver's reaction time, the time needed to apply the brake, etc. These submicroscopic simulation models are very suited to predict the impacts of driver support systems on the vehicle dynamics and driving behavior. Examples of submicroscopic models are SIMONE (Minderhoud, 1995), MIXIC (van Arem and Hogema, 1995), and PELOPS (Ludmann, 1998). For a review on microscopic and submicroscopic simulation models, we refer to Ludmann (1998) and Minderhoud (1995).

Cellular Automaton (CA) or particle hopping models. CA-models aim to combine advantages of micro-simulation models, while remaining computationally efficient by use of efficient storage and computation algorithms. The car-following rules generally lack intuitive appeal and their exact mechanisms are not easily interpretable from the driving-task perspective. These models describe the traffic system as a lattice of cells of equal size (typically 7.5 m). A CA-model describes in a discrete way the movements of vehicles from cell to cell (cf. Nagel, 1996, 1998). The size of the cells are chosen such that a vehicle driving with a speed equal to one, moves to the next downstream cell during one time step. The vehicle's speed can only assume a limited number of discrete values ranging from zero to $v_{\max}$.

The process can be split-up into three steps:

- *Acceleration*. Each vehicle with speed lower than its maximum speed $v_{\max}$, accelerates to a higher speed, i.e., $v \leftarrow \min(v_{\max}, v + 1)$.
- *Deceleration*. If the speed is greater than the distance gap d to the preceding vehicle, then the vehicle will decelerate: $v \leftarrow \min(v, d)$.
- *Dawdling ("Trödeln")*. With given probability p_{brake}, the speed of a vehicle decreases spontaneously: $v \leftarrow \max(v - 1, 0)$.

Using this *minimal set* of driving rules, and the ability to apply parallel computing,[2] the CA-model is very fast, and can consequently be used both to simulate traffic operations on large-scale motorway networks, as well as for traffic assignment and traffic forecasting purposes. The initial single-lane model of Nagel (1996) has been generalized to multilane multiclass traffic flow. In Wu et al. (1999) *time-oriented* car-following rules have been proposed, instead of the traditional space-oriented heuristic rules. It is argued that the resulting model describes drivers' behavior more realistically.

[2] When one relaxes the *parallel update* requirement, we generally do not speak of Cellular Automata models. However, the term *particle hopping model* still applies (cf. Nagel, 1998).

Verification of CA-models for car traffic on German and American motorways and urban traffic networks (Wu et al., 1999; Esser et al., 1999), shows fairly realistic results on a macroscopic scale, especially in the case of urban networks in terms of reproduction of empirical speed-density curves.

Fuzzy logic-based models. The first application of fuzzy logic systems is due to Kikuchi and Chakroborty (1992), aiming at fuzzifying the stimulus-response model. The model was used to illustrate how a fuzzy logic system can be used to describe car-following and local instability. More recent developments are reported in Rekersbrink (1995) and Henn (1995).

2.1.5 Particle models

Particle models can be considered as a specific type of numerical solution approach (so-called *particle discretization methods*; cf. Hockney and Eastwood, 1988), applied to mesoscopic or macroscopic continuum traffic flow models. These models distinguish individual vehicles, but their behavior is described by aggregate equations of motion, for instance a macroscopic traffic flow model. An example of a particle model is INTEGRATION (van Aerde, 1994).

2.1.6 Continuum traffic flow models

Continuum traffic flow models deal with traffic flow in terms of aggregate variables, such as flow, densities and mean speeds. Usually, the models are derived from the analogy between vehicular flow and flow of continuous media (e.g., fluids or gasses), complemented by specific relations describing the average macroscopic properties of traffic flow (e.g., the relation between density and speed). Continuum flow models generally have a limited number of equations that are relatively easy to handle.

Most continuum models describe the dynamics of density $k = k(x, t)$, space mean speed $u = u(x, t)$, and flow $q = q(x, t)$. The density $k(x, t)$ describes the expected number of vehicles per unit length at instant t. The flow $q(x, t)$ equals the expected number of vehicles flowing past cross-section x during per time unit. The speed $u(x, t)$ equals the mean speed of vehicles defined according to $q = ku$. Some macroscopic traffic flow models also contain partial differential equations of the speed variance $\theta = \theta(x, t)$ or the traffic pressure $P = P(x, t) = k\theta$.

Conservation of vehicles and the kinematic wave model. Assuming that the dependent traffic flow variables (density, flow, speed) are differentiable functions of time t and space x, the following partial differential equation represents the fact that on a roadway, vehicles cannot be lost or created:

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = r(x, t) - s(x, t), \tag{11}$$

describing that the number of vehicles on a small part of the roadway of length dx increases according to the balance of inflow and outflow at the boundaries

(interfaces) x and $x + \mathrm{d}x$, respectively, and the inflow $r(x,t)$ and outflow $s(x,t)$ at on-ramps and off-ramps, respectively (source and sink terms). Together with the fundamental relation $q = ku$, Equation (11) constitutes a system of two independent equations and three unknown variables. Consequently, to get a complete description of traffic dynamics, a third independent model equation is needed.

In combining the fundamental relation Equation (1) with Equation (11), a nonlinear first-order partial differential equation results: the *kinematic wave model* (Lighthill and Whitham, 1955):

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = \frac{\partial k}{\partial t} + c(k)\frac{\partial k}{\partial x} = 0, \quad \text{where } c(k) = \frac{\mathrm{d}Q}{\mathrm{d}k}. \tag{12}$$

Here $c(k)$ denotes the so-called kinematic wave speed, describing the speed at which small disturbances propagate through the traffic flow.

Generalized solutions to the kinematic wave model can be determined by the *method of characteristics*, see, e.g., Logghe (2003). For the kinematic wave model, it can be shown that characteristic curves are straight lines in the (x,t)-plane with slope $c(k)$ that emanate from the boundary (i.e., at $x = x_0$ or $t = t_0$) of the considered time–space region. Along the characteristics, densities are conserved and are thus equal to the density at the point on the boundary from which the characteristic emanates. When on this boundary $\partial c(k)/\partial x < 0$, the characteristic curves will in time intersect (*focusing*) and a *shockwave* will result. The shock wave speed ω can be determined from the shock wave equation (May, 1990):

$$\omega = \frac{q_2 - q_1}{k_2 - k_1}, \tag{13}$$

where (k_1, q_1) and (k_2, q_2) respectively denote the traffic conditions downstream and upstream of the shock S. Besides shockwaves, *acceleration fans* are formed in case of discontinuities in the density, characterized by $k(x,t) > k(y,t)$ for $x < y$. These acceleration fans describe the way vehicles drive away from a high-density region into a low-density region. A typical situation in which this occurs, is a traffic light turning to green, where the acceleration fan describes the way vehicles drive away from the formed vehicle queue.

The kinematic wave model can be solved efficiently either analytically or numerically, and its properties and limitations are well understood. Amongst the drawbacks of the model is the formation of shocks irrespective of the smoothness of the initial conditions. Moreover, the kinematic wave model assumes that the traffic speeds adapt to the stationary speed $U(k)$ immediately (no fluctuations around the equilibrium speed), and thus does not respect the finite reaction times and bounded acceleration possibilities of its constituent elements. The latter drawback has been remedied in Lebacque (2002), by imposing additional constraints on the solutions of the kinematic wave model prescribing the maximum acceleration of the cars. The kinematic wave model is not able to predict stop-and-go waves with amplitude-dependent oscillation

times, which are quite common in real-life traffic flow (Verweij, 1985), nor is traffic hysteresis (average headways of vehicles approaching a jam are smaller than vehicles driving out of a jam, see Treiterer and Myers, 1974) described. Traffic instability is also not captured by the kinematic wave model.

Recent improvements in the model are reported in Daganzo (1997, 2002a, 2002b), considering multiple lanes, as well as dividing the driver population into different user-classes showing different driving characteristics. The concept of motivation, indicating that passing drivers will temporarily accept smaller headways, is also introduced.

Payne-type models. To relax the assumptions that the speeds cannot differ from the stationary speed $u = U(k)$, the latter expression has been replaced by a dynamic equation for the speeds alongside the conservation of vehicle equation (Payne, 1971). Payne-like models can be derived from car-following laws. Considering a driver at location $x_i(t)$, *looking ahead to location*

$$x_i^{\varepsilon} = (1-\varepsilon)(x_{i-1} - x_i) + \varepsilon(x_{i-2} - x_{i-1}), \quad \text{where } 0 \leqslant \varepsilon \leqslant 1. \tag{14}$$

In Zhang (2003) the following expression for the speed v_i of vehicle i is used

$$\begin{aligned} v_i(t+T_{\mathrm{r}}) &= U\big(k\big(x_i^{\varepsilon}(t),t\big)\big) + \beta\big(v_i^{\varepsilon}(t) - v_i(t)\big), \\ &\text{where } v_i^{\varepsilon}(t) := u\big(x_i^{\varepsilon}(t),t\big), \end{aligned} \tag{15}$$

U denotes the equilibrium speed as a function of the density, T_{r} denotes the reaction time, and β is a dimensionless parameter. Using Taylor series expansions (Zhang, 2003), the following dynamic expression for the mean speed $v(x,t)$ is derived

$$\frac{\partial u}{\partial t} + \big(u + 2\beta c_*(k)\big)\frac{\partial u}{\partial x} + \frac{c^2(k)}{k}\frac{\partial k}{\partial x} = \frac{U(k)-u}{T_{\mathrm{r}}} + \mu(k)\frac{\partial^2 u}{\partial x^2}, \tag{16}$$

where

$$c_*(k) = k\frac{\mathrm{d}U}{\mathrm{d}k} \quad \text{and} \quad \mu(k) = 2\beta T_{\mathrm{r}} c^2(k) \tag{17}$$

denote the sound speed and the traffic viscosity, respectively.

From Equation (16), different factors can be identified that can be interpreted from driver behavior. The term $(c^2(k)/k)(\partial k/\partial x)$ denotes the effect of driver anticipation, showing how drivers anticipate on downstream conditions: in regions of increasing density, drivers will anticipate and reduce their speeds accordingly. The relaxation term $(U(k)-u)/T_{\mathrm{r}}$ describes the smooth adaptation of the speed u to an equilibrium state $U(k)$, given the relaxation time T_{r}; under stationary conditions, we have $u = U(k)$. The viscosity term $\mu(k)\,\partial^2 u/\partial x^2$ reflects the influence of higher-order anticipation, i.e., the way drivers react to changes in relative speeds $v_i^{\varepsilon}(t) - v_i(t)$.

For specific parameter choices, the general expression (16) can be reduced to other models: the original model of Payne (1971) can be derived by choosing

$\beta = 0$ (no higher-order anticipation). For a constant sound speed $c_*(k) = c_0$, the viscous model of Kühne (1991) results.

It can be shown that the model is *unstable* in a certain density range (small perturbations in the density grow into traffic jams), and that the model is able to (qualitatively) describe stop-and-go traffic. Moreover, small perturbations in the stable flow will dissipate. In general, solutions to the model are smooth. As a result, the model does *backward smoothing* to a sharp concentration/speed profile, thus possibly predicting negative driving speeds. We can therefore conclude that the model may *violate the anisotropic character* (traffic mainly reacts to downstream traffic conditions) of traffic flow. This nonanisotropic nature manifests itself prominently in the workings of shock and expansion waves: in contrast to the kinematic model, which has only one family of kinematic waves, Payne-type models have two, associated with the characteristic curves $\xi_{1,2}$ defined by

$$\frac{\mathrm{d}\xi_{1,2}}{\mathrm{d}t} = \lambda_{1,2} = u + \left(\beta \pm \sqrt{1+\beta^2}\right)c(k). \tag{18}$$

The waves in the first characteristic ξ_1 field travel with a speed less than the speed u, and are qualitatively identical to the kinematic waves in the kinematic wave model. The waves in the second characteristic field travel faster than the average traffic flow, implying that in this field, information reaches vehicles from behind. In Zhang (2003) it is argued that for $\beta < 1$, this will never occur, since the speed of the second characteristic ξ_2 approaches u.

Extensions are reported in Hoogendoorn and Bovy (1999), Hoogendoorn et al. (2002), and Helbing et al. (2001), which pertain to the modeling of multiclass and multilane traffic flow in networks, including nonlocal, forwardly directed interactions, effects of vehicle space requirements. It is important to note that these models are based on gas-kinetic models (see Section 2.1.7), rather than on car-following models like equation (15).

2.1.7 Gas-kinetic flow models

Starting point of the gas-kinetic models, is the so-called phase-space density (PSD),

$$\kappa(x, t, v) = k(x, t) \cdot f(v|x, t), \tag{19}$$

where κ describes the mean number of vehicles $k(x, t)$ per unit roadway length and $f(v)$ the speed distribution at that location and instant. Prigogine and Herman (1971) were the first to use the notion of the PSD to derive a model describing the behavior of traffic flow. They achieved this by assuming that the PSD changes according to the following processes:

1. Convection $\partial(vk)/\partial x$. Vehicles with a speed v flow into and out of the roadway segment $[x, x + \mathrm{d}x)$, causing a change in the PSD $\kappa(x, t, v)$.
2. Acceleration towards the desired speed $(V^0(v) - v)/T_\mathrm{r}$, where $V^0(v)$ denotes the expected desired speed of vehicles driving with speed v; T_r denotes the acceleration time.

3. Deceleration when catching up with a slower vehicle, while not being able to immediately overtake $(1 - p(k))\kappa(x, t, v) \int (w - v)\kappa(x, t, w)\, \mathrm{d}w = (1 - p(k))\kappa(x, t, v)(u(x, t) - v)$.

Their deliberations yielded the following partial differential equation (PH-model):

$$\frac{\partial \kappa}{\partial t} + \frac{\partial (vk)}{\partial x} = \frac{\partial}{\partial v}\left(\frac{V^0(v) - v}{T_\mathrm{r}}\right) + \big(1 - p(k)\big)\kappa(x, t, v)\big(u(x, t) - v\big), \tag{20}$$

where the density k and the mean speed u are defined according to

$$k(x, t) = \int \kappa(x, t, v)\, \mathrm{d}v \quad \text{and}$$
$$u(x, t) = \int v f(v|x, t)\, \mathrm{d}v. \tag{21}$$

The most complex process here is probably the interaction process. Let us briefly discuss how this term is determined from the following, simple behavioral assumptions:

1. The "slow-down event" is instantaneous and occurs with a probability of $(1 - p(k))$, where p denotes the so-called immediate overtaking probability, reflecting the event that a fast car catching up with a slow car can immediately overtake to another lane, without needing to reduce its speed.
2. The speed of the slow car is not affected by the encounter with the fast car, whether the latter is able to overtake or not.
3. The lengths of the vehicles can be neglected.
4. Only two vehicle encounters are to be considered, multivehicle encounters are excluded.

The model of Prigogine and Herman has been criticized and improved by Paveri-Fontana (1975). He considers a hypothetical scenario where a free-flowing vehicle catches up with a slow moving queue, and considers two extreme cases:

1. The incoming vehicle passes the whole queue as if it were one vehicle.
2. It consecutively passes each single car in the queue independently.

In Paveri-Fontana (1975) it is shown that the Prigogine and Herman formalism reflects the second case, while the real-life situation falls between these two extremes. He also shows that the term reflecting the acceleration process yields a desired speed distribution that is dependent on the local number of vehicles. This is in contradiction to the well-accepted hypothesis that driver's personality is indifferent with respect to changing traffic conditions (the so-called *personality condition*; cf. Daganzo, 1995). To remedy this deficiency, Paveri-Fontana

generalized the PSD $\kappa(x, t, v)$ by also including the distribution of the desired speeds

$$\tilde{\kappa}(x, t, v, v_0) = k(x, t)\tilde{f}(v, v_0|x, t), \tag{22}$$

where $\tilde{f}(v, v_0|x, t)$ denotes the joint probability density function of speed v and free speed v_0.

Other researchers have objected to the validity of the *vehicular chaos assumption* underlying the expression for the effects of vehicle interactions. In Munjal and Pahl (1969) it is argued that the interaction term "corresponds to an approximation in which correlation between nearby drivers is neglected", being only valid in situations where no vehicles are platooning. This issue has been remedied explicitly in Hoogendoorn and Bovy (1999) by distinguishing between platooning and nonplatooning vehicles.

In Nelson et al. (1997) it is argued that plausible speed-density relations can only be determined from the Prigogine–Herman model, based on the nontrivial assumption that the underlying distribution of desired speeds is *nonzero for very small speeds*. The situation when this assumption does not hold is investigated in Nelson and Sopasakis (1998). It is found that at concentrations above some critical value, there is a two-parameter family of solutions, and hence a continuum of mean velocities for each concentration. This result holds for both constant values of the passing probability and the relaxation time, and for values that depend on concentration in the manner assumed by Prigogine and Herman. It is hypothesized that this result reflects the well-known tendency toward substantial scatter in observational data of traffic flow at high concentrations.

Paveri-Fontana model generalizations are reported in Hoogendoorn and Bovy (1999), Hoogendoorn et al. (2002), and Helbing et al. (2001), where gas-kinetic models for multiclass and multilane traffic flow including nonlocal, forwardly directed interactions, effects of vehicle space requirements are presented. These gas-kinetic models serve as the starting point to derive continuum models by application of the so-called *method-of-moments*. Another multilane gas-kinetic model was proposed in Klar and Wegener (1999a, 1999b). In Tampére et al. (2002) adaptive driver behavior is introduced into the gas-kinetic modeling approach.

2.1.8 Model application

Traffic flow and microsimulation models designed to characterize the behavior of the complex traffic flow system have become an essential tool in traffic flow analysis and experimentation. The application areas of these tools are very broad, e.g.:

- Evaluation of alternative treatments in (dynamic) traffic management.
- Design and testing of new transportation facilities (e.g., geometric designs).

- Operational flow models serving as a submodule in other tools (e.g., traffic state estimation, model-based traffic control and optimization, and dynamic traffic assignment).
- Training of traffic operators.

Which or even *if* a model should be used, depends largely on the type of problem at hand. Important issues are the purpose of the study, the required level-of-detail (is the individual driver behavior and changes therein important), what kind of data is available for model calibration and validation, and the type of network considered (urban, motorway).

Nevertheless, some general remarks can be made. For one, the application of microscopic (simulation) models will in general be more time-consuming, both in the sense of computation time needed to perform the simulations (long computation time per simulation due to detailed representation of dynamic processes) and requirement to do multiple runs to get statistically valid results in case of stochastic microsimulation models.

Moreover, calibration and validation of microscopic models may be a laborious task. This can be explained by noticing that these models aim to mimic human behavior in real-life traffic (not in contrived "car-following experiments"), which is hard to observe, measure and validate (cf. Daganzo, 1994). This is problematic, given the observed nonlinear behavior of the collective traffic flow: the microscopic details have to be just right for the simulation to realistically describe and predict for instance stop–start waves in traffic flow. In Brackstone and McDonald (1999) it is convincingly argued that suitable data (e.g., pair-wise vehicle trajectories collected by instrumented vehicles, or remote-sensing; cf. Hoogendoorn et al., 2002) must be used in the model calibration, and that the models are to be disassembled and tested in a step-by-step fashion. In general, the lack of "microscopic" data results in macroscopic calibration that cannot produce the optimal parameters since the number of degrees of freedom is too large.

Macroscopic models are generally suited for large scale, network-wide applications, where macroscopic characteristics of the flow are of prime interest. Macroscopic models are generally too coarse to correctly describe microscopic details and impacts, for instance caused by changes in roadway geometry. Macroscopic models are assumed to describe macroscopic characteristics of traffic flow more accurately. Calibration of macroscopic models is *relatively* simple (compared to microscopic models), for instance using loop detector data (see Cremer and Papageorgiou, 1981; Helbing, 1997). Mostly, speed-density relations derived from observations are required. In a recent paper, Kerner et al. (2000), it is shown that traffic jam dynamics can be described and predicted using macroscopic models that feature only some characteristic variables, which are to a large extent independent on roadway geometry, weather, etc. This implies that macroscopic models can describe jam propagation reliably, without the need for in-depth model calibration.

3 Route guidance and information systems

3.1 Introduction

Advances in data processing, sensor and communications technologies have made it possible to provide travelers with information on network conditions based on real-time measurements. Better information should enable individuals to make better travel decisions. Moreover, as significant numbers of travelers respond to such information, network conditions will themselves be affected. Of particular interest here are systems intended to improve route choice decisions, either by providing data on network conditions or by recommending specific routes to a destination. These are called Route Guidance and Information Systems (RGIS). A number of factors justify interest in these systems.

To begin with, travelers are frequently unaware of all the options available to them. This is clearly the case for those unfamiliar with an area; the increasing popularity of GPS-based navigation systems that provide turn-by-turn directions to a destination, attests to the value of basic way-finding information. In addition, there is considerable evidence that even individuals who consider themselves familiar with an area, have only a limited knowledge of travel options. In Jeffery (1981), for example, it is estimated that, with better information on travel options and conditions, even habitual drivers in an area could reduce their average distance and travel time by around 7 percent.

Moreover, congestion is to a significant extent an unpredictable phenomenon. It has been estimated, for example, that roughly 60 percent of congestion delays on US urban highways result from specific unpredictable events such as accidents, vehicle breakdowns, and the like (Lindley, 1987). Recurrent congestion, resulting from traffic levels that are systematically high relative to available roadway capacity over a particular time period, also has a random component that derives from variability in travel demand levels and network performance. Because of the randomness in travel conditions, prior experience can be an imperfect basis for travel decisions, and supplementing it with more up-to-date information could result in better individual decisions as well as, perhaps, improved network conditions.

A number of researchers have assessed the likely network-level impacts and benefits of RGIS. Notable studies include Koutsopoulos and Lotan (1989), Mahmassani and Jayakrishnan (1991), Al-Deek and Kanafani (1993), Emmerink et al. (1995), and Hall (1996). Under a variety of assumptions and approaches, these studies have investigated the likely reduction in total travel time, the distribution of travel time savings between guided and unguided drivers, and the variation of these benefits as a function of the RGIS market penetration rate (fraction of drivers receiving guidance).

On the other hand, operational experiences with RGIS to date have generally been on too limited a scale and for too brief a time to allow strong

empirical conclusions about its network level impacts to be drawn. Possible exceptions include high-volume corridors equipped with variable message signs, and urban areas with real-time traffic condition reports. For example, using data from the Washington, DC traffic information system (Wunderlich et al., 2001) travelers with and without access to information on prevailing link traversal times were simulated. Travelers were assumed to have a desired arrival time at their destination, and to determine their path and departure time accordingly. The simulated travel experiences, in terms of travel times, on-time arrival reliability, risk of lateness, and early and late schedule delays were compiled. It was found that guidance improved the various measures of travel time reliability without significantly affecting average travel time itself.

This section reviews current knowledge about route guidance and information systems. Although a distinction is sometimes made between prescriptive *guidance* and descriptive *information*, both kinds of data will generally be referred to here as guidance. A particular set of guidance data disseminated at a particular time will be called a *message*. Objectives and technological constraints that influence message content are discussed in Section 3.2.

Messages may be derived from static or dynamic information about the network. Static systems provide fixed information about the network and may be of use in tasks such as way-finding or preliminary trip planning; however, they do not recognize actual traffic conditions. Static systems will not be further considered here. Dynamic RGIS can be classified as nonpredictive and predictive systems. The former base the messages provided to drivers on measurements or estimates of prevailing network conditions, while the latter derive messages from forecasts of future network states. The two kinds of system can involve quite different issues, and will be discussed separately in Sections 3.3 and 3.4.

Data on the effects of guidance on individual traveler behavior are available from laboratory experiments with driving simulators and, to a more limited extent, from surveys and observations of travelers who use RGIS. Knowledge of these effects is important to develop predictive guidance, and ultimately to evaluate the economic benefits of RGIS. Current knowledge of traveler response to information is discussed in Section 3.5.

Finally, Section 3.6 identifies some areas of current research.

3.2 Overview of route guidance and information systems

3.2.1 Guidance objectives

In traffic networks subject to congestion, a flow pattern that optimizes a system-level objective is not generally the same as one that results when each driver independently chooses her preferred route. For example, the flow pattern that minimizes the total travel time of all vehicles on the network generally differs from the pattern obtained when each driver attempts to minimize her individual travel time. The two types of pattern are referred to as system and user optimal flow patterns, respectively.

It is sometimes suggested that network operators could use route guidance as a tool to force the network towards a system optimal flow pattern. This may be an appropriate policy in exceptional circumstances such as emergency evacuation situations, but it is unlikely to be successful in the long run for routine situations. Drivers are free to ignore any guidance messages that they perceive as incompatible with their own decision criteria. Moreover, systematic attempts to influence drivers' decisions by providing misleading information, even if well intentioned, are likely to lead over time to large scale driver rejection of the RGIS.

The focus here will be on guidance derived from user optimality objectives. As argued in Hall (1996), such guidance can be usefully viewed as a way of correcting the dis-equilibrium travel behavior that results from lack of information. However, it may not be feasible to provide individual drivers with messages precisely matched to their particular route choice decision criteria. A common alternative approach is to base messages solely on travel times, i.e., to report link or path traversal times (descriptive information) or to recommend minimum travel time paths (prescriptive guidance). Traversal times prevailing at the time of guidance generation are often used for this purpose; these are called *instantaneous* travel times. However, in a dynamic network, with link traversal times that vary over time, instantaneous times may be different from the *experienced* times that drivers incur when making a trip. While guidance based on experienced times is arguably closer to drivers' own choice criteria, generating such guidance requires the use of a predictive model, and is considerably more complex than simply measuring and reporting the prevailing traversal times.

Guidance is an attempt to improve the information available to drivers for their route choice decision, yet the guidance itself will rarely be perfect. Data collection, processing, and communications systems constrain the quality and quantity of data that can be generated and transmitted, and humans are limited in their ability to process information, in particular while driving. These factors affect the type of guidance that an RGIS can provide. A poorly designed RGIS can exacerbate rather than improve congestion. In Ben-Akiva et al. (1991) and Kaysi et al. (1995) RGIS design issues are discussed including possible counterproductive effects of poorly-designed systems. These include *concentration*, in which guidance reduces the normal dispersion of driver behaviors and leads to increased congestion on a smaller number of routes; and *overreaction*, in which drivers' response to guidance shifts congestion or leads to oscillations in flows on different routes.

3.2.2 RGIS functional characteristics

The principal features that characterize different RGIS designs are briefly identified here. Although the features are presented separately, many of them are, in fact, interrelated in the sense that a choice for one constrains the feasible options for others.

Basis for guidance. As was indicated above, a major distinction is between nonpredictive systems that generate guidance based on measurements or estimates of prevailing network conditions, and predictive systems that utilize forecasts for this purpose.

Local/area-wide focus. Guidance generation may focus on either a local or a wide area traffic network. A local focus considers conditions on individual or perhaps small contiguous groups of network elements (road segments or junctions); the guidance will typically be disseminated only over that area. A wide area focus considers conditions throughout the network in determining the messages to disseminate.

Transmission range. The RGIS dissemination technology determines the distance from a guidance source over which the messages may be received. Line-of-sight, small area and wide area technologies are possible. In the case of short- and medium-range technologies, the locating of the guidance dissemination sources, and the resulting coverage of the network and its flows, are important system design decisions.

Collective/individual dissemination. The RGIS dissemination technology also determines whether messages can be received by all vehicles in transmission range, or only by vehicles equipped with suitable receivers. Examples of the former include roadside variable message signs and highway advisory broadcasts over standard radio frequencies; examples of the latter include coded infrared, microwave, FM sub-carrier, and cellular packet radio transmissions.

One-way/two-way communications. In an one-way communication system, drivers receive messages from the RGIS but do not provide any information to the system. In a two-way system, drivers notify the system about their travel desires and receive messages that are tailored to their specific trip needs. Moreover, the data acquired about travel times and drivers' trip intentions and choices can be incorporated in the system's state estimates and forecasts.

Pre-trip/en-route access to guidance. Guidance received prior to beginning a trip may influence the decision to travel or not, the destination(s), the time of departure, the mode of travel as well as the particular route to follow. Guidance received while en-route will generally only affect the subsequent choice of path. Driver response to messages may also be different in the two situations: a pre-trip decision is a choice without immediate antecedent, whereas the en-route decision to switch routes may involve a reluctance to abandon a prior choice and so exhibit hysteresis or a threshold effect.

Message dissemination and guidance update intervals. Guidance messages generally relate to a period of time rather than to a single instant, and so are maintained or retransmitted during that period. The message dissemination interval is the period of time during which a disseminated guidance message does not change. The length of this period may be dictated by technological constraints or by human factors. The guidance update interval refers to the time between successive computations of the messages to disseminate. In a complex guidance system, messages cannot be continually recomputed because of

delays inherent in collecting and processing data, and in generating and disseminating the messages.

Message design. The final issue concerns the syntax and semantics of the guidance messages themselves, including their medium of delivery, format, content, and precision. The distinction between prescriptive guidance and descriptive information was mentioned above. Visual or audio messages intended for direct reception by drivers cannot be overly complex because of the difficulty of assimilating them while driving. Messages that will be processed by an in-vehicle unit and conveyed to the driver in a schematic visual form might perhaps have a higher data content. The available communications bandwidth or message display capabilities may also constrain message complexity and precision.

3.3 Nonpredictive and related systems

In a nonpredictive dynamic RGIS, messages are derived from estimates of the network conditions that prevail at or before the time the guidance is generated. To the extent that these instantaneous conditions are a good indication of what a driver will actually encounter during a trip, then route choice decisions made using nonpredictive guidance should be well founded. Conversely, the potential limitation of nonpredictive guidance is that network conditions may change significantly during a trip and so invalidate decisions based solely on conditions around the time of departure.

Descriptive nonpredictive guidance consists of information on estimated prevailing traffic conditions, by processing measurements collected from various kinds of traffic sensors. Details of such data collection and processing methods are beyond the scope of this paper. Prescriptive nonpredictive guidance consists of route recommendations based on estimated prevailing conditions. Numerous methods have been proposed for generating such guidance. In Papageorgiou (1990), Bolelli et al. (1991), and Charbonnier et al. (1991), for example, nonpredictive route guidance approaches based on methods from control engineering have been proposed.

Most currently operational dynamic RGIS are nonpredictive. They typically disseminate messages using variable message signs, public radio and television, or telephones. In-vehicle dynamic nonpredictive RGIS have mostly been limited-scale experimental prototypes. For example, the TravTek system (Rilett et al., 1991), deployed in Orlando, Florida, included an in-vehicle unit that received by radio coded updates of prevailing link travel times and computed the minimum time path to destinations selected by the driver.

3.4 Predictive systems

3.4.1 Predictive guidance and consistency

Predictive route guidance and information systems base guidance on forecasts of future network conditions. The messages disseminated to a driver

reflect expectations of what conditions will be at network locations at the time the driver will actually be there, and so are arguably closer to driver's actual decision criteria than nonpredictive messages based on prevailing or historical conditions. In Ben-Akiva et al. (1996) the conditions under which nonpredictive and predictive guidance result in the same flow patterns were analyzed. In Pavlis and Papageorgiou (1999) it was showed that, in densely meshed networks, nonpredictive guidance can have the same effects as predictive guidance. In general, however, nonpredictive and predictive guidance will be different and will have different impacts on network conditions.

Predictive guidance messages are derived from forecasts of future network conditions. However, when these messages are disseminated and drivers react to them in some way (for example, by changing departure times or paths), future conditions are likely to be affected, possibly invalidating the forecasts and rendering the messages irrelevant or worse. Within the context of a model system, predictive guidance is said to be *consistent* if the forecasts on which it is based are the same, within the limits of modeling accuracy, as those predicted to result after drivers receive the guidance and react to it. Consistency is a generalization of the concept of traffic equilibrium. Unlike conventional equilibrium, which assumes that travelers are perfectly informed about network conditions, consistency accounts for the specific characteristics of available travel information and driver response to it.

3.4.2 Approaches for predictive guidance generation

One approach to generating predictive guidance simply ignores the consistency issue. Guidance messages are based on condition forecasts that are extrapolations from prevailing and historical conditions. The limitation of this approach is that it does not take account of the effects of the guidance itself on future network conditions: it does not ensure guidance consistency. This may not be important at low levels of driver participation in the RGIS, but is likely to lead to incorrect guidance messages when the number of vehicles responding to guidance is sufficient to affect network flows and conditions.

Computation of consistent predictive route guidance requires application of a dynamic traffic network model in order to forecast network conditions. Such models are high-dimensional nonlinear systems; they can be quite complex and require considerable amounts of data to identify. (Again, the necessary data collection and processing tasks are not considered here.) They take the form of mathematical models when analytical tractability is important, and of simulation models when the involved relationships defy convenient mathematical expression.

Such a model is driven by exogenous time-dependent origin–destination (OD) demands. The model includes components that predict how these demands will choose paths and propagate along them. (Some models also predict travelers' choice of departure times.) The propagation of demand over its chosen paths is determined by time-dependent link traversal times, and these

traversal times are in turn affected by the congestion that results from the movement of demand along the links.

Conventional (nonguidance) dynamic traffic assignment models compute time-dependent equilibrium flows and traversal times under the assumption that demand has perfect information about present and future congestion levels and chooses paths accordingly. Guidance-oriented traffic network models, on the other hand, must explicitly consider the availability and nature of travel information, as well as driver behavior in the presence and absence of such information. In Ben-Akiva et al. (1991) and Watling and van Vuren (1993) the features that dynamic network models require for route guidance applications are considered.

Guidance-oriented traffic models have been less studied than conventional traffic assignment models. A high-level formal representation of a network model for predictive guidance generation can be obtained using three time-dependent problem variables and three maps (that are implemented as models and algorithms) that relate them. The variables are: C, the network conditions; M, the guidance messages; and P, the path splits (fraction of trips going to a particular destination via each available path or subpath) at trip origins and en-route decision points. The maps are:

- the *network loading map* $S : P \rightarrow C$, which determines the network conditions that result from the movement of exogenous time-dependent OD demands over the network in accordance with a particular set of path splits;
- the *routing map* $D : M \rightarrow P$, which determines the path splits that result from a particular set of guidance messages. The routing map generally incorporates a model of driver response to guidance messages; and
- the *guidance map* $G : C \rightarrow M$, which represents the response of the RGIS, in the form of guidance messages, to a given set of network conditions. (Note that messages output by this map for a given set of conditions are not necessarily consistent, since driver reaction to the messages may result in network conditions different from the inputs.)

Composite problem maps can be obtained by combining the network loading, routing and guidance maps in different sequences. Each composite map transforms an element of one problem variable into another element of the same variable. There are three such maps (the symbol $*$ denotes functional composition):

- a composite map $D * G * S : P \rightarrow P$ from the domain of path splits into itself;
- a composite map $S * D * G : C \rightarrow C$ from the domain of link conditions into itself; and
- a composite map $G * S * D : M \rightarrow M$ from the domain of messages into itself.

In terms of these composite problem maps, predictive guidance consistency means that a map's time-dependent inputs (i.e., the time trajectory of a problem variable) coincide with its time-dependent outputs. Equality of the map's input and output values means that the value is a *fixed point* of the map. (If $T: X \to X$ is a one-to-one map, $x^* \in X$ is a fixed point of T if $x^* = T(x^*)$.) Thus, consistency in the context of predictive route guidance can be computed and studied in terms of fixed points of the problem maps, or of approximations to them.

Fixed point approaches under perfect information assumptions have been investigated in Kaufman et al. (1998) as a basis for dynamic traffic assignment and in Engelson (1997) in the context of driver information systems. In Yang (1998) and Bovy (1999) guidance models using fixed point approaches were also considered. In Bottom (2000) additional details on the predictive guidance generation formulation presented here are provided.

3.4.3 Operational predictive guidance systems

Operational experience with predictive systems is currently very limited.

The LISB system in Berlin (Hoffman, 1991) and the Autoguide system in London (Catling, 1989) were early prototypes of systems with in-vehicle and infrastructure-based components. Communications between the two components served to establish travel times as well as to disseminate guidance. The guidance consisted of next turn recommendations derived from minimum path calculations using simple link traversal time predictions. These predictions, in turn, were derived from historical link traversal time patterns and recently measured traversal times. The usage of these systems did not attain levels that would create consistency issues.

In the Netherlands, dynamic route information panels (DRIPs) can display route recommendations for simple network topologies based on traffic predictions and optimal control laws (Hoogendoorn and Bovy, 1998).

At least one traffic data company in the US sells network condition predictions that are based on extrapolations that take account of current traffic measurements, typical patterns of link condition variations over time, and other factors such as weather and special events. The details of the extrapolation method are considered secret.

Systems that provide consistent predictive guidance at the level of an urban or regional network are not yet operational. There have been limited experimental deployments in traffic control centers of software systems (for example, DYNASMART-X and DynaMIT) that are designed to generate such guidance. The guidance generation logic in DynaMIT is explicitly based on the fixed point approach described in the preceding section.

3.5 Driver response to RGIS

Understanding driver response to RGIS is required to develop effective guidance systems that meet drivers' information needs and contribute to con-

gestion relief. Predictive guidance, in particular, requires the ability to forecast driver responses to different possible messages in order to ensure guidance consistency. Moreover, the economic evaluation of guidance systems also requires a knowledge of the range of traveler responses to travel information.

Data on driver response to RGIS come from laboratory experiments with travel choice simulators and, to a lesser extent, from observations of traveler interactions with operational systems. The following paragraphs briefly summarize what is currently known and not known regarding the impacts of information on various travel-related decisions.

RGIS awareness and access decisions. Awareness, willingness to pay, and usage rates can be obtained by conducting travel surveys in areas where RGIS services are available (Polydoropoulou and Ben-Akiva, 1999). The large-scale panel surveys conducted every 3 years in Seattle-area by the Puget Sound Regional Council provide data on the evolution over time of awareness and usage decisions for different RGIS types (Peirce and Lappin, 2002).

Decision to travel or not. Relatively little information is available regarding the effects of RGIS on the decision to travel or not; however, it is not inconceivable that information about sufficiently bad travel conditions could induce travelers to cancel their intended trips, particularly discretionary trips. In Khattak et al. (1999) evidence are cited for this effect among noncommuters from surveys carried out as part of the San Francisco-area TravInfo project.

Choice of destination or destinations. Similarly, relatively little information is available in the literature regarding the effects of RGIS on destination choice, or on the decision to visit several destinations and accomplish several purposes in one trip through trip-chaining. Trips offering a choice among multiple destination alternatives are likely to be for shopping (see, for example, Kraan et al., 2000) or personal purposes. Opportunities to group multiple purposes and destinations into a single trip-chain are more varied and difficult to characterize.

Departure time choice. In Mannering et al. (1994) results from Seattle-area surveys about commuters who receive travel information from radio, television, and telephone services are presented. Of the commuters surveyed, 40% indicated that they had some flexibility in scheduling and selecting the route for their morning commute trip; 23% indicated no flexibility. However, 64% responded that they rarely changed their departure time because of pre-trip information.

Mode choice. Little detailed information is available about the mode choice impacts of RGIS, although there is some evidence for this effect. As reported in Yim and Miller (2000), less than 1% of the early callers to San Francisco's TravInfo service asked to be rerouted to the transit menu after learning about bad traffic conditions from the traffic menu. However, as experience with the system increased over the duration of the TravInfo field test deployment, it was found that up to 5% of the callers asked to be rerouted to the transit menu. Of those who accessed transit information, 90% ultimately chose transit for their travel mode.

Route choice. Many surveys and travel choice simulator studies have demonstrated the ability of RGIS to influence route choice. Based on analysis of driver route choice responses to both VMS and radio information, it has been suggested in Emmerink et al. (1996) that some people have an innate propensity to use traffic information of any kind and from any source. Nonetheless, there is considerable evidence that the nature of the guidance information, and the conditions experienced prior to its dissemination, can strongly affect driver route choice response to it.

Drivers' perceptions of the accuracy and reliability of the messages is a key determinant of their response. It has been found (Kantowitz et al., 1997) that there exists an accuracy "threshold", beneath which drivers will simply ignore RGIS messages. Factors that increase drivers' confidence in the accuracy of the messages tend to increase the likelihood that the drivers will react to them. In the context of route choice, such factors include observation of congestion prior (and particularly just prior) to receiving the message, and favorable experiences with the RGIS in prior uses. Drivers appear to be tolerant of a certain amount of error in RGIS messages, although drivers familiar with an area will expect a higher degree of accuracy from the information system.

Some drivers express a strong preference for descriptive information on traffic conditions, while others prefer prescriptive recommendations of a particular route to take (Khattak et al., 1996; Polydoropoulou et al., 1996). Combining a prescriptive recommendation to change routes with descriptive information justifying the recommendation has been found in some travel choice simulator experiments to result in the highest route switching compliance rates (Bonsall and Palmer, 1999).

A number of generally idiosyncratic factors condition a driver's route choice response to RGIS messages. For example, a motorway bias has been observed in several studies. Because of this bias, drivers receiving messages that suggest diverting from a nonmotorway to a motorway facility are considerably more likely to comply than those who receive the opposite message, other things being equal. As mentioned above, habit also plays a significant role in travel decisions.

Learning. The day-to-day dynamics of commuter pre-trip departure time and route choices as well as en-route path switching for morning commutes were analyzed in Mahmassani and Liu (1999). Factors affecting route choice behavior include: (1) arrival time flexibility, (2) user characteristics, and (3) information reliability. In Ozbay et al. (2001) the use of a stochastic learning algorithm to analyze drivers' day-to-day route choice behavior is proposed. This model addresses the learning behavior of travelers based on experienced travel time and day-to-day learning.

3.6 Areas of current research

Further development of route guidance and information systems will require better understanding of a number of issues, many of which have only

recently begun to receive attention. This section describes ongoing research in RGIS architecture, real-time computing, stochasticity and driver behavior.

System architecture. Nonpredictive guidance systems are relatively simple in conception and robust in operation. Although they use data on instantaneous network conditions, these systems may sometimes succeed in attaining objectives based on experienced conditions, but need not do so in general. Basing guidance on instantaneous conditions may sometimes exacerbate rather than improve traffic problems. Predictive systems, based on experienced conditions, depend on the availability and reliability of complex models of traveler behavior and network performance. Furthermore, they may be sensitive to high levels of noise in model predictions, and are computationally demanding. Is there a guidance system architecture that combines the better features of the two approaches while avoiding their drawbacks? For example, multilevel control system designs have been developed for traffic control systems, but have been less investigated in the context of RGIS.

Real-time response. Predictive guidance generation for a realistic network requires considerable amounts of computation, yet it must be done quickly and accurately enough for the guidance to be timely and of use to drivers. Parallel and distributed computation environments are of interest in this regard, as are fixed-point solution heuristics.

Stochasticity. Any of the individual maps involved in the predictive guidance generation problem may be stochastic. The composite problem map will then be stochastic as well, and its output when evaluated will be a realization of a stochastic process rather than the deterministic time trajectory of a problem variable. Nonetheless, the fixed-point interpretation of guidance consistency continues to apply in this case, with the understanding that consistency now means that problem map inputs and outputs are both stochastically equivalent realizations of the same stochastic process. Markov chain Monte Carlo techniques such as Gibbs sampling may be used to compute problem variable statistics to any desired degree of accuracy. However, this approach is very computationally demanding.

In practice, most stochastic guidance modeling efforts have adopted a "noisy map" approximation to address the effects of stochasticity in the computation of guidance solutions. Implicitly or explicitly, these approaches treat model outputs not as a realization of a general stochastic process but rather as a time trajectory of deterministic values affected by noise. Stochastic approximation procedures are then applied to compute the fixed point. No rigorous justification has yet been provided for this approach.

Driver behavior modeling. Applications of RGIS require the development of reliable models of driver behavior and, in particular, of their response to guidance messages. An important aspect of this is the development of better models of the ways in which travelers form new perceptions from their most recent experience, the guidance they received and their earlier experiences. These efforts will benefit from advances in the understanding of the psychological and cognitive processes involved in decision-making. Of particular interest

are studies of decision-making under the time pressure of driving situations, and studies of the ways in which spatial and network knowledge affect driver response to RGIS.

4 Urban network traffic control

4.1 Introduction

Optimum management and control of traffic in urban networks is an important requirement for city authorities as they seek efficient, safe and sustainable transport. In addition there is an increasingly wide range of demanding objectives for transport policy makers to achieve, such as public transport priority, improved conditions for vulnerable road users, real-time traffic information; emergency and incident management and restraining traffic in sensitive areas. As a response to these issues, Urban Traffic Management and Control (UTMC) systems have been introduced in many cities around the world to provide the tools to support efficient and effective network management to meet needs of current and future traffic problems. Fundamentally UTMC systems are conceived as modular, open systems that incorporate and build on existing functionalities of existing signal control and other traffic management systems as illustrated in Figure 3. An important point to note is that the Urban Traffic Control (UTC) systems are often at the heart of UTMC and provide a better migration path so that improvements in UTC are utilized to the full in UTMC.

UTC refers to the control of traffic in urban areas using traffic signals, which are linked to operate in a coordinated way. Such linked signal systems may be used to achieve a variety of policy objectives, which relate to efficiency of traffic operations, improved safety, reduced atmospheric pollution, priority for specific road user groups, access control to maintain or enhance urban environments, and to mitigate the effects of irregular events such as accidents or road closure. UTC systems use historic or (more commonly) real-time knowledge of network conditions to determine the control strategy most appropriate for the conditions, and signal infrastructure to inform and control road users.

4.2 UTC systems: general requirements

Early systems in the 1950s and 1960s were based on fixed-time traffic control providing signal coordination or progression for traffic on an arterial, through the optimization of offsets between adjacent sets of signals. UTC was therefore justified on there being a sufficient density of traffic signals to make signal co-ordination worthwhile, compared to the alternative of operating traffic signals in isolation. Whilst relatively effective for traffic co-ordination in "predictable" conditions, the inability of fixed-time systems to adjust to changing traffic conditions has been a drawback in this approach. The desire for traffic signaling

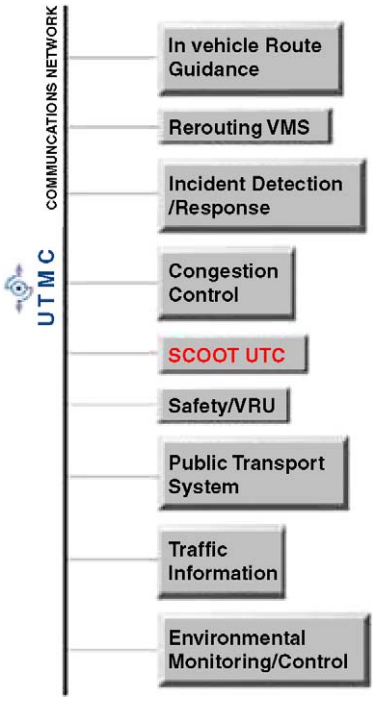


Fig. 3. Schematic illustration of a UTMC system (Source: Department of the UK Environment, Transport and the Regions (1999))

to be more responsive to changing traffic conditions has led to the development of a range of semi or fully traffic responsive UTC systems. The improved performance of these systems has generally justified their additional cost (e.g., detection, maintenance, etc.).

A variety of methods for UTC have evolved over the last decades, responding to the needs of individual cities/countries, the existing research and development base and advances in detection, communications and control technology. These traffic-responsive UTC systems are continuously upgraded to meet with current requirements. Quality attributes of a UTC system play a major role in its architecture. These may include attributes such as the speed of system response to recurring congestion and incidents (i.e., responsiveness), feedback philosophy, ability of integration, functional and spatial extendability, wider range of control strategies, robustness, installation and maintenance costs, etc. Flexibility of the system to incorporate enhancements as policies/technologies advance is a further key attribute.

Criteria for installing a real-time UTC system are now much wider than the need for efficient signal coordination for traffic. For example, the UTC communications infrastructure and processing capabilities give a powerful tool for

the network manager, including such functions as traffic information, automatic incident detection (AID), and congestion management.

4.3 Fundamentals of UTC

4.3.1 Principles of traffic signal control

Traffic signals operate by giving sequential priority to movements, including pedestrian/cycling stages and other priorities. Sufficient separation of stages is essential to ensure that conflicts between movements do not occur and there are a variety of regulations and guidance documents provided by government agencies. The regulations are similar across many countries, and differences reflect local driver behavior/expectation, enforcement regimes, and national attitudes to guidance, regulation, and safety. In general, the smaller the amount of time lost to road users when changing signal priority, the greater the capacity and hence the shorter the delay.

Traffic signal controlled intersections may operate in isolation or be linked to one or more adjacent intersections as part of a coordinated approach, i.e., UTC. An isolated traffic signal is usually set by estimating an optimum cycle time and green splits (the green times allocated to each separate movement), i.e., those which minimize the delay. The optimum times are determined from a knowledge of traffic demand and the maximum (or "saturation") flows, also taking account of the time lost to traffic movements when the signals are red to all traffic and the time lost as flows build up and fall from maximum discharge rates. Usually delay is minimized when the degree of saturation (i.e., the ratio of demand to capacity, for the key movement) is about 0.85. Capacity is a key parameter affecting performance and is determined as the saturation flow multiplied by the effective green time per hour available for that flow. When linking signals, the offset (i.e., the time delay in the start of the downstream cycle) is the crucial third variable to be considered with green splits and cycle time. The need for linking signals usually relates to their proximity to each other and the extent to which linking allows "platoons" of traffic to proceed through adjacent junctions more efficiently.

For any specific condition of traffic demand a range of algorithms may be used to optimize signal settings. In the early UTC systems, the database of traffic demand was assumed fixed and signal timings were determined off-line. However, traffic demand exhibits substantial short-term variability as well as longer term changes in levels of flow and movement patterns. This has led to increasingly sophisticated approaches to UTC which rely on substantial detector input. Broadly, UTC systems may be categorized as fixed-time using historic databases or demand responsive using on-line traffic data inputs. The latter may be subdivided into centralized or decentralized systems. The characteristics of these systems are outlined below in Sections 4.3.2 and 4.3.3 and further details of specific systems are given in Section 4.4.

4.3.2 Fixed-time systems

In fixed-time systems, off-line optimization is undertaken using demand levels which are assumed constant for the period over which each fixed-time plan is intended to run. Up to 10–15 plans may be developed to represent the complete set of traffic conditions to be found on the network at different times. A network is considered to operate as a series of different regions of groups of linked signals, within each of which different signal plans will run. To ensure that the arrival patterns of successive platoons of vehicles arrive consistently at the downstream signals:

(1) a single cycle time must apply across the region and
(2) time offsets of the starts of successive cycles at one intersection from the next must be the same.

The regional cycle time is based on that required for the busiest intersection. Thus, a region may be bounded either by road links along which the benefits of linking are small or where a common cycle time is very inappropriate.

Fixed-time plans may be readily used to create green waves, give predetermined priorities, and respond to special events which can be predicted, such as football matches. They cannot respond to unplanned incidents such as traffic accidents or unplanned road works. Plans may be set to change at predetermined times or changes may be triggered by flow or queue measurements taken at key locations. There are also several systems which generate new plans on-line, i.e., using very recent historic data.

Fixed-time systems require a considerable amount of traffic data to be collected to set up and to keep up-to-date. Fixed-time plans can age rapidly, particularly where traffic growth is high, and the benefits of linking may be lost in three to four years if the plans are not updated. A further problem occurs when plans change and discontinuities in flow patterns occur. This limits the number of plans which can be used. Both the above points can be addressed using traffic-responsive systems.

4.3.3 Traffic-responsive systems

Traffic-responsive systems use on-line detector measurements to optimize signal timings on a cycle-to-cycle basis to better meet demand. Such systems may be coordinated largely from a central computer, e.g., SCOOT (Bretherton, 1998) or have distributed intelligence and be coordinated largely at a local level, e.g., UTOPIA (Donati et al., 1984). Centrally controlled systems use less intelligent local controllers, whilst with decentralized systems each controller is more capable of taking local decisions, with some coordination between adjacent controllers. A wide range of traffic responsive systems are now available with varying degrees of central and local control and key systems are described in Section 4.4.

If a system is to respond to changes in traffic conditions, comprehensive detection must be available. Detectors must be accurate, located appropriately for the characteristics of the UTC system, and the information must be reliably sent to the appropriate control center. In general, the more sophisticated

the system the more comprehensive the detector requirements and the more susceptible it is to detector failure. Many systems have default values for the controllers based on time of day which are implemented if loss of detectors or communication occurs.

Increasingly, a wide range of detectors are available. Traditionally, ground-based systems using inductive loops to measure the presence of a vehicle have formed the basis of most UTC detection. Other ground-based systems include magnetometers which measure changes in the earth's magnetic field brought about by the presence of a vehicle. Above ground detectors include microwave systems, radar, infra-red, video, and laser systems. Each has specific characteristics to capture different aspects of vehicle behavior. Image-based systems can be installed without costly and disruptive installation works, but have yet to reach their full potential. Using vehicles themselves as detectors is an application being considered for the future. Overall, the quality, quantity, and reliability of future information will encourage more sophisticated UTC control strategies.

Table 2 provides a summary of the main advantages and disadvantages of different types of UTC systems.

4.4 System summaries

4.4.1 Fixed-time systems

TRANSYT (TRAffic Network StudY Tool). TRANSYT (Robertson, 1997) is the most well-developed and widely-used fixed time UTC system. It is an off-line program for calculating optimum coordinated signal timings in a network of traffic signals. For each distinct traffic stream it assumes that the flow rate averaged over a specified period is known and constant and that the saturation flows for each link are also known. TRANSYT consists of two main elements called the "traffic model" and the "signal optimizer", as shown in Figure 4. The traffic model represents traffic behavior in a highway network and predicts a performance index (PI) for the network for a given fixed-time plan and average set of flows on each link. The PI measures the overall cost of traffic "congestion", which is usually a weighted combination of the total delay and the number of stops made by vehicles.

Cyclic Flow Profiles (CFP's) showing the distribution of flows entering each link are used with a "platoon dispersion" model to estimate patterns of vehicle arrivals at the downstream junction. "Uniform" delay is calculated in a similar way as for SCOOT, illustrated in Figure 5, supplemented by formulae to represent random delays and oversaturated delays when the junction is over-loaded (i.e., the queue does not clear in the green period). Signal optimization involves an iterative "hill climbing" process to adjust the signal timings to achieve an optimum PI. Specific links may be given extra weighting by the user to implement, for example, green waves on a corridor. Other city/country specific fixed-time UTC systems are used around the world, but are not as widespread nor as well documented as TRANSYT, so they are not described further here.

Table 2.
Summary of advantages and disadvantages of different types of UTC systems

UTC system	Advantages	Disadvantages
Fixed-time	1. Cheaper to install and maintain. 2. Can be implemented using noncentrally controlled equipment. 3. Familiarity with settings for regular users. 4. Green waves more easily implemented. 5. Can favor specific vehicle types easily.	1. Large amount of data to be collected and updated. 2. Signal plans may require updating. 3. Disruption of plan changing. 4. Operator reaction to incidents required. 5. Can not deal with short-term traffic fluctuations.
Responsive plan selection	1. Can deal with some day to day fluctuations. 2. Plan change time could be more appropriate. 3. Might be valuable on arterial routes. 4. Cheaper than fully responsive systems.	1. Requires more data than fixed-time systems. 2. Detector failures possible. 3. Needs discussions on thresholds for plan change. 4. Plan may change for a wrong reason. 5. Difficult to foresee all plan needs.
Fully responsive	1. Less data needed to be collected in advance. 2. Plan evolves, so avoids problems with plan changing and updating. 3. Can deal with short and long term traffic fluctuations. 4. Automatic reaction to incidents. 5. Monitors traffic situation throughout the area.	1. Detector failures possible. 2. More expensive to install and maintain. 3. Requires some central control. 4. Maintenance critical.

4.4.2 Traffic-responsive systems: Centralized

With centralized control, traffic detector information is sent to the UTC center where it is processed and "optimum" timings are calculated for all the traffic signals within the UTC system. These timings are then sent back to each traffic signal controller on-street. Intelligence is therefore retained at one location (the UTC center). The costs of on-street controllers can then be less, although communication costs will usually be higher than decentralized systems with distributed intelligence. Five such systems are summarized below.

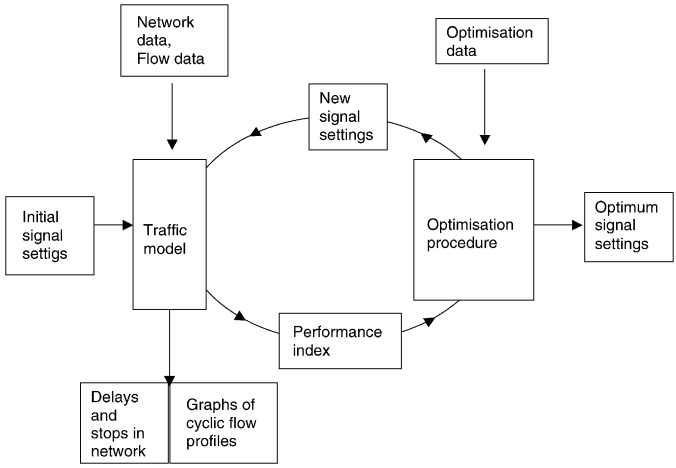


Fig. 4. Traffic model and signal optimizer in TRANSYT.

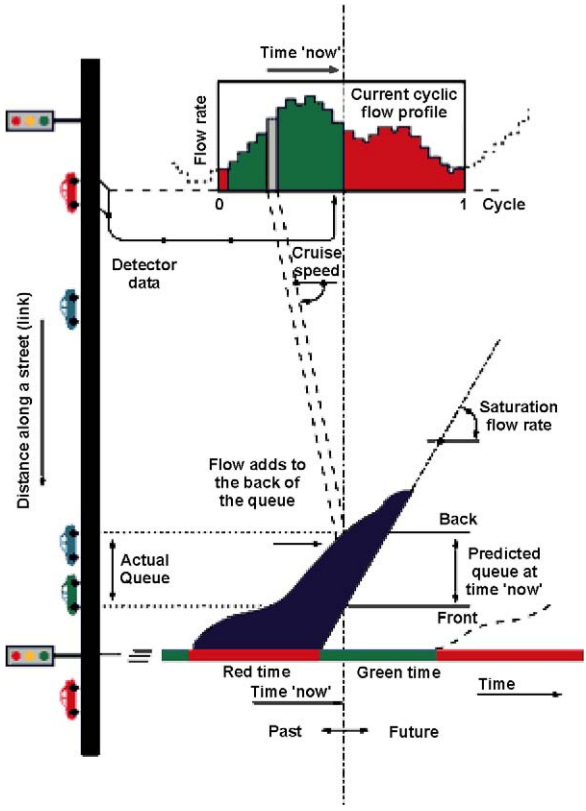


Fig. 5. Principles of the SCOOT traffic model (Source: Department of the UK Environment, Transport and the Regions (1999)).

SCOOT (Split Cycle Offset Optimization Technique). SCOOT (Hunt et al., 1981; Bretherton, 1998) was developed in the United Kingdom and is operational in many cities around the world. It operates based on three main principles, namely the measurement of cyclic flow profiles (CFP), the updating every 4 seconds of its on-line traffic model of queues and delays on each link and incremental optimization of signal timings. SCOOT uses detectors at the upstream end of links to measure demand and CFP's in real time. The upstream detection also allows any congestion on the link to be monitored (i.e., when the queue reaches the upstream detector) and the possible exit blocking effects of this congestion on upstream links. The SCOOT model predicts downstream arrival patterns using a calibrated link cruise speed with some dispersion. The saturation flow rate for each signal stop line is validated when the system is commissioned; this allows the growth and clearance of queues to be estimated accurately. The on-line traffic model is used in real time by the signal optimizer. SCOOT has three optimization procedures by which it adjusts signal timings (Department of the UK Environment, Transport and the Regions, 1999). These are the cycle time, green splits, and offsets, each optimized using a different procedure at different frequencies. By the combination of relatively small changes to traffic signal timings, SCOOT can respond to both short-term local peaks in traffic demand, as well as following trends over time and thus maintain a constant coordination of the signal network.

In addition to the optimization from the basic SCOOT model, the operation has considerable flexibility to override values and set parameters for different regions and different times. These may include gating strategies to protect an area from excessive levels of traffic, bus priorities, etc. In addition to network management, SCOOT has a substantial data base facility for storing, manipulating, and presenting traffic data including flows, journey times, and queues. Further facilities have been included in the latest releases (Bretherton et al., 2003) with further research ongoing (Bretherton et al., 2004).

SCATS (Sydney Coordinated Adaptive Traffic System). SCATS (Lowrie, 1982) was developed in Australia and has been implemented in many cities around the world. It operates at two basic levels known as the "upper" level, which involves offset plan selection and the "lower" level, which involves optimization of junction parameters. The upper level generates offset plans by time of day from historic data while the lower junction level optimizes green splits, cycle times, and offsets between signalized junctions using an incremental feedback process based largely on detectors situated at the stop lines. SCATS calculates green splits based on the flow in the previous cycle and so is not fully responsive to unpredictable arrival flows. It differs from systems such as SCOOT and UTOPIA in that it does not have a traffic model and uses stop line detectors to estimate departure rates, rather than arrival rates modeled from upstream detectors.

SCATS is basically a modular system largely run by regional computers capable of handling a large number of intersections, with significant intelligence

within local controllers. A central computer may also be used to improve management functions. SCATS differs from many other systems in that the network manager has a more direct involvement in setting up the system, i.e., it does not have a model. The degree of operator understanding increases with the level of simplicity of a system and this would lead to corridor operations being addressed most beneficially.

RHODES (Real-Time Hierarchical Optimized Distributed and Effective System). RHODES (Head et al., 1992) was developed in United States. The RHODES architecture is based on three levels of hierarchy. The highest level assigns traffic to the network to determine base levels of traffic across the network which takes into account both evolving traffic demand and current network geometry. At the next level down, RHODES operates as a more typical UTC system based on predicted platoon arrival patterns. At the intersection level the movements of individual vehicles are modeled.

Basically there are two processes in RHODES namely "estimation and prediction" and "decision system". The first process takes the upstream detector data and estimates the actual flow profiles in the network and the subsequent propagation of these flows. On the other hand, the second process is where the phase durations are selected to optimize a given objective function (minimization of average delay per vehicle, average queue lengths, numbers of stops, etc.), the optimization being based on dynamic programming and decision trees. Recently RHODES has been updated for the integration of bus priority measures (Mirchandani et al., 2001).

MOTION (Method for the Optimization of Traffic Signals in Online Controlled Networks). MOTION (Busch, 1996) is a recent UTC system developed in Germany, with some limited implementation in some European cities. MOTION is basically based on four functional levels namely "data acquisition and pre-processing", "traffic modeling and analysis", "optimization of control variables", and "decision and transfer of signal programs". The first module receives the dynamic information from detection equipment via the central traffic computer and may perform some data processing functions like determination of origins and destinations. The second module uses most important individual traffic streams to determine actual O–D streams within the network and turning movements at intersections, which is necessary to calculate green splits and minimum cycle times for each intersection. Selection of a common network cycle time for coordination, determination of link progression speeds, and optimal offsets between individual intersections to minimize delay and stops are determined in the third module. On the fourth level new signal programs are evaluated and transferred via the central traffic computer.

TUC (Traffic-Responsive Urban Control). TUC (Diakaki et al., 2000) has been developed recently in Greece and has also been implemented in a few other European cities, particularly in the context of EC-funded demonstration

projects. TUC operates by modifying nominal signal timings using a multivariable regulator on-line. Nominal starting values for signal timings are based on historic levels of demand. The regulator is based on the formulation of the urban traffic control operation as a linear–quadratic control problem. The control objective is to minimize and balance link queues taking into account link storage capacity. This formulation is potentially particularly useful in dealing with oversaturated conditions. TUC has been expanded in recent years to allow for public transport priority (Diakaki et al., 2003).

4.4.3 Traffic-responsive systems: decentralized

With decentralized (distributed) control, more intelligence for signal optimization is distributed to local traffic signal controllers. This can increase flexibility and reduce communications costs, but controllers are usually more costly. Three distributed systems are summarized below.

UTOPIA (Urban Traffic Optimization by Integrated Automation). UTOPIA (Donati et al., 1984) was originally developed in Italy. The system is structured as a hierarchical system organized on three levels known as the local level, the area level and the town supervisor level. UTOPIA's intelligent local controllers can communicate with each other as well as with a central computer. The local outstation level applies a microscopic model to estimate the state of the intersection directly collecting data from detectors located at the start of each link. Local queue and turning percentage estimation, saturation flows and delay calculations are performed by the local "observer". The next level uses a historic traffic database to validate the local detection, checking changes in the traffic data or making comparisons of data upstream and downstream of the congested links. The final level integrates the congestion information with data from other systems like public transportation. The macroscopic model used at this level has the advantage of collecting different sources of information and having the coverage of the whole city.

PRODYN. PRODYN (Farges, 1990) was developed in France and has been implemented in some other European cities. It uses an intersection open-loop optimal feedback algorithm for traffic signal control. As with SCOOT and UTOPIA, detectors are located at the upstream end of each link and where appropriate at 200 m and 50 m upstream. The detectors collect occupancy data. The system operates in 5 sec steps and the demand for each period is estimated from that in the previous period. A time horizon for prediction is 75 sec. Optimization seeks to minimize the sum of the delays over the horizon. A forward dynamic programming procedure is used for optimization. Intersection controllers simulate the outputs over the horizon using the link outputs and off-line determined turning proportions. Intersection controllers communicate with each other to achieve a better arrival forecast for the downstream intersection. The control structure at the network level is a decentralized one.

OPAC (Optimization Policies for Adaptive Control). OPAC (Gartner, 1991) was first developed in the United States using dynamic programming to generate optimal control strategies. It provides the computation of signal timing without requiring fixed cycle time, split, and offset in the conventional sense, and it is constrained only by minimum and maximum green times. OPAC calculates, in real time, near-optimal signal timings using on-line data that is typically readily available from upstream detectors at local level and OPAC supports system-wide coordination at the network level. Many developments have been carried out over the years (Valdes and Paz, 2004).

4.4.4 Performance

A variety of studies have been undertaken in different locations seeking to compare the performance of alternative control systems. Early comparisons were between isolated and coordinated forms of control. Results would be expected to be highly dependent on network characteristics, so that, co-ordination should be most favorable on arterial routes with closely spaced traffic signals. Probably the most detailed surveys were undertaken in Glasgow, where fixed time co-ordination was found to reduce vehicle journey times by some 16% on average compared to isolated control (Holroyd and Hillier, 1979).

Further comparisons by the UK Transport Research Laboratory have found that the SCOOT UTC system offers delay savings of around 12% compared to up-to-date fixed-time plans and up to 40% in peak periods in networks operating under isolated vehicle actuated control (McDonald and Hounsell, 1991). A 4% annual increase in delay has also been reported for fixed-time plans if not updated (Bell and Bretherton, 1986), so that the potential benefits of traffic-responsive systems would then be higher.

Performance of the other systems described in Sections 4.4.1 and 4.4.2 have also generally been evaluated through "before-and-after" studies. For example, surveys of UTOPIA in Turin gave reductions in journey times of 20% for public transport vehicles and 10–15% for other vehicles. Good results have also been reported for SCATS and PRODYN. However, there is very little evidence of the comparative performance of the systems described on the same network.

4.5 Discussion

4.5.1 Operational Research (OR) techniques

The significant increase in real-time information availability on traffic states in recent years, driven by advances in technology for detection, communications, and data processing, opens up exciting opportunities for further OR applications. It is beyond the scope of this paper to discuss these in detail, but opportunities are evident even from a sample of OR-related techniques already being used, such as:

- Short-term prediction/forecasting (e.g., evolution of traffic states).

- Data fusion (data increasingly available from different sources).
- Closed and open-loop control theory applications.
- Dynamic programming for optimization.
- Advanced control theory applications.
- Prediction methods for platoon dispersion, time-dependent queuing, etc.
- Real-time simulation modeling and network analysis.
- Optimization methods (e.g., for signal timings to optimize against increasingly diverse objective functions).
- Applications of fuzzy-logic for modeling/optimization.
- Artificial intelligence and "expert" systems.

Perhaps a key challenge in the coming years will be how to select and use OR techniques most effectively against a background of changing optimization criteria and data provision which, whilst rich, will be inevitably variable in quantity, quality, and coverage.

4.5.2 Concluding comments

The optimum use of traffic signals for urban network traffic management and control will continue to be a key issue for City Authorities. This section therefore concludes with some comments on some of the opportunities and challenges which can be identified for the coming years.

- Accurate and timely data will remain a key requirement of UTC systems: In general, the less timely the data, the poorer the control (Bretherton et al., 2004). However, this can impose a considerable cost burden on detection/communications, and advantage is yet to be fully taken of new above-ground systems. Good real-time modeling of congested situations remains an important component of effective UTC (Jhaveri et al., 2003). This would seem to be a priority area for scientific research, given the increase in congestion occurring in many towns and cities.
- UTC systems will increasingly have to provide flexibility in control strategy selection, including priorities for public transport, pedestrians, and other road user groups. This will have an implication for optimization methods and criteria.
- Integration of UTC with other physical and ITS-related urban traffic management systems can offer significant benefits for the network and will be a key requirement for the coming years.

5 Motorway traffic control

5.1 The control loop

Controlling the motorway traffic flow process is a highly complicated task which may involve a variety of spatially distributed control measures such as

ramp metering, route guidance, variable speed limits, etc. The way the control measures behave and act on the traffic process stems from the specific design of the control strategy used. The control strategy employed determines the control actions, and the specific response to the prevailing traffic conditions, through the available control actuators, is based on its design and on pre-specified goals.

Figure 6 depicts the general control loop for the motorway network traffic process which includes all technical and physical phenomena that should be influenced according to the specific goals. The evolution of the traffic process depends upon the control inputs and the process disturbances. The control inputs are directly related to corresponding control devices such as traffic lights, variable message signs, variable direction signs, etc., and may be selected from an admissible control region subject to technical, physical, and operational constraints. The process disturbances cannot be manipulated, but may possibly be measurable (e.g., demand) or detectable (e.g., incident) or predictable over a future time horizon with appropriate algorithms. Typical disturbances in motorway traffic are traffic demands, origin–destination patterns, the drivers' compliance to variable message signs, environmental conditions, and incidents.

The process outputs are quantities chosen to represent the performance aspects of interest, e.g., total time spent, queue lengths, etc. The estimation of the traffic state and the prediction of the various traffic quantities are performed based on real-time measurements taken from the traffic process, and are subsequently fed to the control strategy. The control strategy determines, based

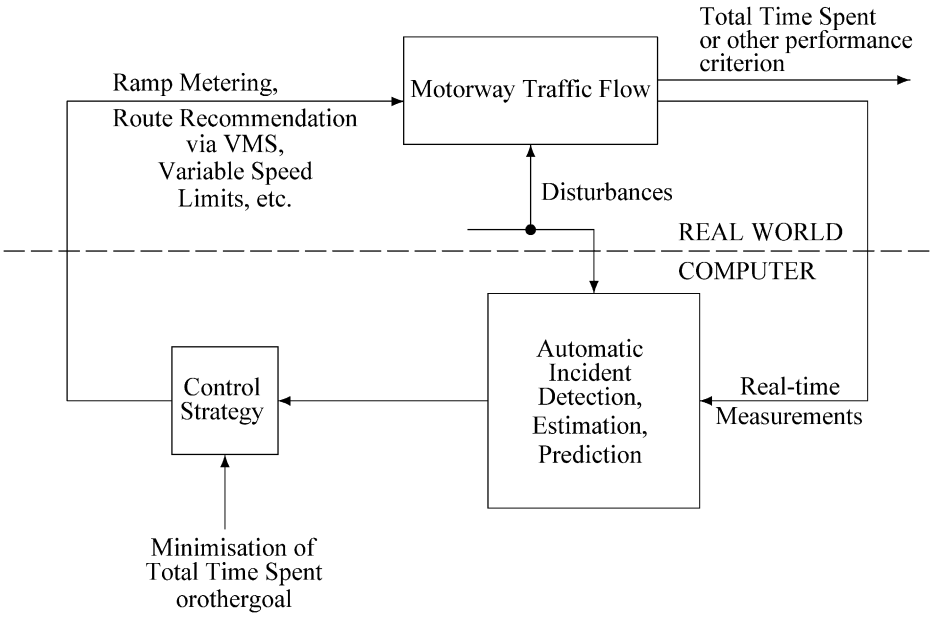


Fig. 6. Motorway traffic flow process under control.

on the measured, estimated, and predicted quantities, the appropriate control inputs which are fed to the traffic process so as to meet the specified goals despite the impact of various disturbances.

5.2 Control strategy design

5.2.1 Ramp metering control strategies

General remarks. Ramp metering is the most direct and efficient way to control and upgrade motorway traffic. Various positive effects are achievable if ramp metering is appropriately applied:

- Increase in mainline throughput due to avoidance or reduction of congestion.
- Increase in the served volume due to avoidance of blocked off-ramps or motorway interchanges.
- Utilization of possible reserve capacity on parallel arterials.
- Improved traffic safety due to reduced congestion and safer merging.

Fixed-time ramp metering strategies. Fixed-time ramp metering strategies are derived off-line for particular times-of-day, based on constant historical demands, without the use of real-time measurements. They are usually based on simple static models (Wattleworth, 1965; Schwartz and Tan, 1977).

As an objective criterion, one may wish to maximize the number of served vehicles (which is equivalent to minimizing the total time spent), or to maximize the total travel distance, or to balance the ramp queues. These formulations lead to linear programming or quadratic programming problems that may be readily solved by use of broadly available computer codes. An extension of these methods that renders the static model dynamic by introduction of constant travel times for each section was suggested in Papageorgiou (1980).

The main drawback of fixed-time ramp metering strategies is that their settings are based on historical rather than real-time data. This may be a rude simplification because:

- Demands are not constant, even within a time-of-day.
- Demands may vary at different days, e.g., due to special events.
- Demands change in the long term leading to "aging" of the optimized settings.
- Turning movements are also changing in the same ways as demands; in addition, turning movements may change due to the drivers' response to the new optimized signal settings, whereby they try to minimize their individual travel times.
- Incidents and further disturbances may perturb traffic conditions in a nonpredictable way.

In addition, fixed-time ramp metering strategies may lead (due to the absence of real-time measurements) either to overload of the mainstream flow (congestion) or to underutilization of the motorway.

5.2.2 Reactive ramp metering strategies

Reactive ramp metering strategies are employed at a tactical level, i.e., in the aim of keeping the motorway traffic conditions close to pre-specified set values, based on real-time measurements.

Local ramp metering. Local ramp metering strategies make use of traffic measurements in the vicinity of a ramp to calculate suitable ramp metering values. The demand-capacity strategy (Masher et al., 1975), quite popular in North America, reads

$$r(k) = \begin{cases} q_{\text{cap}} - q_{\text{in}}(k-1) & \text{if } o_{\text{out}}(k) \leqslant o_{\text{cr}}, \\ r_{\min} & \text{else}, \end{cases} \tag{23}$$

where (Figure 7) k is the discrete time index, q_{cap} is the motorway capacity downstream of the ramp, q_{in} is the motorway flow measurement upstream of the ramp, o_{out} is the motorway occupancy measurement downstream of the ramp, o_{cr} is the critical occupancy (at which the motorway flow becomes maximum), and $r_{\min}$ is a pre-specified minimum ramp flow value. The strategy (23) attempts to add to the measured upstream flow $q_{\text{in}}(k-1)$ as much ramp flow $r(k)$ as necessary to reach the downstream motorway capacity q_{cap}. If, however, for some reason, the downstream measured occupancy $o_{\text{out}}(k)$ becomes overcritical (i.e., a congestion may form), the ramp flow $r(k)$ is reduced to the minimum admissible flow $r_{\min}$ to avoid or to dissolve the congestion.

Comparing the control problem in hand with Figure 6, it becomes clear that the ramp flow r is a control input, the downstream occupancy o_{out} is an output, while the upstream motorway flow q_{in} is a disturbance. Hence, (23) does not really represent a closed-loop strategy but an open-loop disturbance-rejection policy (Figure 7(a)) which is generally known to be quite sensitive to various further nonmeasurable disturbances.

The occupancy strategy (Masher et al., 1975) is based on the same philosophy as the demand-capacity strategy, but it relies on occupancy-based estimation of q_{in}, which may, under certain conditions, reduce the corresponding implementation cost.

An alternative, closed-loop ramp metering strategy (ALINEA) (Figure 7(b)), suggested in Papageorgiou et al. (1991), reads

$$r(k) = r(k-1) + K_{\text{R}}\big[\hat{o} - o_{\text{out}}(k)\big], \tag{24}$$

where $K_{\text{R}} > 0$ is a regulator parameter and $\hat{o}$ is a set (desired) value for the downstream occupancy (typically, but not necessarily, $\hat{o} = o_{\text{cr}}$ may be set, in which case the downstream motorway flow becomes close to q_{cap}). In field applications, ALINEA has not been very sensitive to the choice of the regulator parameter K_{R}.

Note that the demand-capacity strategy reacts to excessive occupancies o_{out} only after a threshold value (o_{cr}) is exceeded, and in a rather crude way, while ALINEA reacts smoothly even to slight differences $\hat{o} - o_{\text{out}}(k)$, and thus it may prevent congestion by stabilizing the traffic flow at a high throughput level.

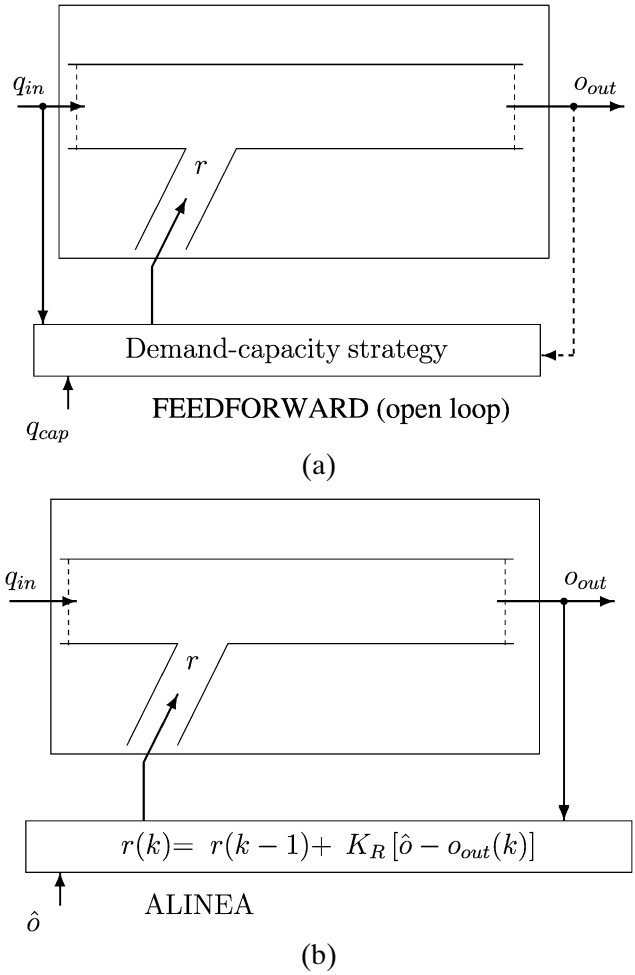


Fig. 7. Local ramp metering strategies. (a) Demand–capacity, (b) ALINEA.

The set value may be changed any time, and thus ALINEA may be embedded into a hierarchical control system with set values of the individual ramps being specified in real time by a superior coordination level or by an operator.

Comparative field trials have been conducted in various countries to assess and compare the efficiency of local ramp metering strategies (see, e.g., Papageorgiou et al., 1998), such as the demand-capacity, ALINEA, and the occupancy strategy. The field results clearly show ALINEA's superiority for all employed performance criterions.

Multivariable regulator strategies. Multivariable regulators for ramp metering pursue the same goals as local ramp metering strategies: they attempt to op-

erate the motorway traffic conditions near some pre-specified set (desired) values. While local ramp metering is performed independently for each ramp, based on local measurements, multivariable regulators make use of all available mainstream measurements $o_i(k)$, $i = 1, \ldots, n$, on a motorway stretch, to calculate simultaneously the ramp volume values $r_i(k)$, $i = 1, \ldots, m$, for all controllable ramps included in the same stretch (Papageorgiou et al., 1990). This provides potential improvements over local ramp metering because of more comprehensive information provision and because of coordinated control actions. Multivariable regulator approaches to ramp metering have been reported in Yuan and Kreer (1968), Young et al. (1997), and Benmohamed and Meerkov (1994). The multivariable regulator strategy METALINE may be viewed as a generalization and extension of ALINEA, whereby the metered on-ramp volumes are calculated from

$$\mathbf{r}(k) = \mathbf{r}(k-1) - \mathbf{K}_1\big[\mathbf{o}(k) - \mathbf{o}(k-1)\big] + \mathbf{K}_2\big[\widehat{\mathbf{O}} - \mathbf{O}(k)\big], \tag{25}$$

where $\mathbf{r} = [r_1, \ldots, r_m]^\top$ is the vector of m controllable on-ramp volumes, $\mathbf{o} = [o_1, \ldots, o_n]^\top$ is the vector of n measured occupancies on the motorway stretch, $\mathbf{O} = [O_1, \ldots, O_m]^\top$ is a subset of $\mathbf{o}$ that includes m occupancy locations for which pre-specified set values $\widehat{\mathbf{O}} = [\widehat{O}_1, \ldots, \widehat{O}_m]^\top$ may be given. Note that for control-theoretic reasons the number of set-valued occupancies cannot be higher than the number of controlled on-ramps. Typically one bottleneck location downstream of each controlled on-ramp is selected for inclusion in the vector $\mathbf{O}$. Finally, $\mathbf{K}_1$ and $\mathbf{K}_2$ are the regulator's constant gain matrices that must be suitably designed via an LQ procedure, see Papageorgiou et al. (1990), and Diakaki and Papageorgiou (1994), for details.

Nonlinear optimal ramp metering strategies. Reactive ramp metering strategies may be helpful to a certain extent, but, first, they need appropriate set values, and, second, their character is more or less local. What is needed for motorway networks or long stretches is a superior coordination level that calculates in real time optimal set values from a proactive, strategic point of view. Such an optimal control strategy should explicitly take into account:

- Demand predictions over a sufficiently long time horizon.
- The current traffic state both on the motorway and on the on-ramps.
- The limited storage capacity of the on-ramps.
- The ramp metering constraints regarding maximum queues allowed.
- The nonlinear traffic flow dynamics, including the infrastructure's limited capacity.
- Any incidents currently present in the motorway network.

Based on this comprehensive information, the control strategy should deliver set values for the overall motorway network over a future time horizon so as

- to respect all present constraints,
- to minimize an objective criterion such as the total time spent in the whole network including the on-ramps.

Such a comprehensive dynamic optimal control problem may be formulated and solved with moderate computation time by use of suitable solution algorithms. The nonlinear traffic dynamics may be expressed by use of suitable dynamic models in state space form, where the state vector comprises all traffic densities and mean speeds of motorway segments, as well as all ramp queues; the control vector comprises all controllable ramp volumes; the disturbance vector comprises all on-ramp demands and turning rates at bifurcations. The problem's constraints include the ramp metering constraints and the queue constraints, see Section 5.3.

Thus, for given current (initial) state from corresponding measurements and given demand predictions, the problem consists in specifying the ramp flows $\mathbf{r}(k)$, $k = 0, \ldots, K-1$, where K is the considered horizon, so as to minimize the total time spent (or some other criterion) subject to the nonlinear traffic flow dynamics and the constraints, see Section 5.3.

This problem or variations thereof was considered and solved in various works (Blinkin, 1976; Kotsialos et al., 2002). Although simulation studies indicate substantial savings of travel time and substantial increase of throughput, advanced control strategies of this kind have not been implemented in the field as yet. Section 5.3 contains a simulation study for such an advanced coordinated ramp metering control strategy.

5.2.3 Link control strategies

Link control may include one or a combination of the following actions:

- Variable speed limitation.
- Changeable message signs with indications for "keep lane", or congestion warning, or environmental warning (e.g., information about the pavement state).
- Lane control.
- Incident warning.
- Reversable flow lanes (tidal flow).

There are many motorway stretches, particularly in Germany and in the Netherlands, employing a selection of these measures. It is generally thought that control measures of this kind lead to a homogenization of traffic flow (i.e., more homogeneous speeds of cars within a lane and of average speeds of different lanes) which is believed to reduce the risk of falling into congestion at high traffic densities and to increase the motorway's capacity. Very few systematic studies have been conducted to quantify the impact of these control measures (see, e.g., Zackor, 1972; Smoulders, 1990) and corresponding validated mathematical models are currently lacking. This is one of the reasons why the corresponding control strategies of operating systems are of a heuristic character (e.g., Bode and Haller, 1983; Zackor and Balz, 1984).

5.2.4 Route guidance control strategies

A route guidance system may be viewed as a traffic control system in the sense of Figure 6. Based on real-time measurements, sufficiently interpreted

and extended within the surveillance block, a control strategy decides about the routes to be recommended (or the information to be provided) to the road users. This, on its turn, has an impact on the traffic flow conditions in the network, and this impact is reflected in the performance indices. Because of the real-time nature of the operation, requirements of short computation times are relatively strict (for more details see Section 3).

5.2.5 *Integrated motorway network traffic control*

As mentioned earlier, modern motorway networks may include different types of control measures. The corresponding control strategies are usually designed and implemented independently, thus failing to exploit the synergistic effects that might result from coordination of the respective control actions. An advanced concept for integrated motorway network control results from suitable extension of the optimal control approach outlined above. More precisely, the dynamic model of motorway traffic flow may be extended to enable the inclusion of further control measures, beyond the ramp metering rates $\mathbf{r}(k)$. Formally $\mathbf{r}(k)$ is then replaced by a general control input vector $\mathbf{u}(k)$ that comprises all implemented control measures of any type. Such an approach was implemented in the integrated motorway network control tool AMOC (Advanced Motorway Optimal Control) in Kotsialos et al. (1999), where ramp metering and route guidance are considered simultaneously with promising results, see also Moreno-Banos et al. (1993), Ataslar and Iftar (1998), Bellemans (2003), Hegyi et al. (2003), and Hegyi (2004).

5.3 *An advanced example*

5.3.1 *The motorway network traffic model*

The efficiency and the amelioration potential of nonlinear optimal ramp metering strategies may be demonstrated by means of simulation for a large-scale network with the use of the AMOC generic motorway network control tool. In this case AMOC does not consider routing control measures, but only ramp metering control actions.

The network is represented by a directed graph whereby the links of the graph represent motorway stretches. Each motorway stretch has uniform characteristics, i.e., no on-/off-ramps and no major changes in geometry. The nodes of the graph are placed at locations where a major change in road geometry occurs, as well as at junctions, on-ramps, and off-ramps.

The time and space arguments are discretized. The discrete-time step is denoted by T. A motorway link m is divided into N_m segments of equal length L_m. Each segment i of link m at time instant $t = kT$, $k = 0, \ldots K$, is characterized by the macroscopic variables traffic density (see also Section 2) $\rho_{m,i}(k)$ (veh/lane-km), mean speed $v_{m,i}(k)$ (km/h), and traffic volume or flow $q_{m,i}(k)$ (veh/h). The basic equations used for their calculation for each segment i of link m at each time step, are (this is a time-space discretized

Payne-like model, see Section 2.1.6)

$$\rho_{m,i}(k+1) = \rho_{m,i}(k) + \frac{T[q_{m,i-1}(k) - q_{m,i}(k)]}{L_m \lambda_m}, \tag{26}$$

$$q_{m,i}(k) = \rho_{m,i}(k) v_{m,i}(k) \lambda_m, \tag{27}$$

$$\begin{aligned} v_{m,i}(k+1) = v_{m,i}(k) &+ \frac{T\{V[\rho_{m,i}(k)] - v_{m,i}(k)\}}{\tau} \\ &+ \frac{T[v_{m,i-1}(k) - v_{m,i}(k)]v_{m,i}(k)}{L_m} \\ &- \frac{\nu T}{\tau L_m} \frac{\rho_{m,i+1}(k) - \rho_{m,i}(k)}{\rho_{m,i}(k) + \kappa}, \end{aligned} \tag{28}$$

$$V\big[\rho_{m,i}(k)\big] = v_{f,m} \exp\left[-\frac{1}{a_m}\left(\frac{\rho_{m,i}(k)}{\rho_{\mathrm{cr},m}}\right)^{a_m}\right], \tag{29}$$

where $v_{f,m}$ denotes the free-flow speed of link m, $\rho_{\mathrm{cr},m}$ denotes the critical density per lane of link m (the density where the maximum flow in the link occurs), λ_m its number of lanes, and a_m is a parameter of the fundamental diagram (Equation (29)) of link m. Furthermore, τ, a time constant, ν, an anticipation constant, and κ, are constant parameters same for all network links. Additionally, it is assumed that the mean speed resulting from (27) is limited from below by the minimum speed in the network $v_{\min}$.

In order for the speed calculation to take into account the speed decrease caused by merging phenomena and the speed reduction due to weaving phenomena, resulting from lane drops in the mainstream, two additional terms are added to (27), see Messmer and Papageorgiou (1990).

For origin links, i.e., links that receive traffic demand and forward it into the motorway network, a simple queue model is used.

$$w_o(k+1) = w_o(k) + T\big[d_o(k) - q_o(k)\big], \tag{30}$$

where $w_o(k)$ is the queue length (veh) in origin o during period k, $d_o(k)$ is the demand (veh/h) at o at the same period, and $q_o(k)$ is the flow (veh/h) that enters the mainstream. The outflow $q_o(k)$ is determined by the traffic conditions on the mainstream link and possible ramp metering control measures applied. If ramp metering is applied, then the outflow $\hat{q}_o(k)$ that is allowed to leave o during period k is a portion $p_o(k)$ of the outflow that would leave o without control.

$$q_o(k) = p_o(k)\hat{q}_o(k), \tag{31}$$

where $p_o(k) \in [p_{\min,o}, 1]$ is the metering rate for the origin link o, i.e., a control variable. If $p_o(k) = 1$, no ramp metering is applied, else $p_o(k) < 1$. For $\hat{q}_o(k)$ we have

$$\hat{q}_o(k) = \min\big\{\hat{q}_{o,1}(k), \hat{q}_{o,2}(k)\big\}, \tag{32}$$

with

$$\hat{q}_{o,1} = d_o(k) + \frac{w_o(k)}{T}, \tag{33}$$

$$\hat{q}_{o,2} = \begin{cases} Q_o & \text{if } \rho_\mu(k) < \rho_{\mathrm{cr},\mu}, \\ Q_o\big[1 - \frac{\rho_{\mu,1}(k) - \rho_{\mathrm{cr},\mu}}{\rho_{\max} - \rho_{\mathrm{cr},\mu}}\big] & \text{if } \rho_\mu(k) \geqslant \rho_{\mathrm{cr},\mu}, \end{cases} \tag{34}$$

where Q_o is the on-ramp's capacity (veh/h), and $\rho_{\max}$ (veh/lane-km) is the maximum density in the network. Thus the maximum outflow $\hat{q}_o(k)$ is determined by the current origin demand if $\hat{q}_{o,1} < \hat{q}_{o,2}$ (see (32), (33)), or the geometrical ramp capacity Q_o if the mainstream density is undercritical, i.e., $\rho_{\mu,1}(k) < \rho_{\mathrm{cr},\mu}$ (see (34)), or the reduced capacity due to congestion of the mainstream, i.e., $\rho_{\mu,1}(k) > \rho_{\mathrm{cr},\mu}$ (see (34)).

Motorway bifurcations and junctions (including on-ramps and off-ramps) are represented by nodes. Traffic enters a node n through a number of input links and is distributed to the output links according to

$$Q_n(k) = \sum_{\mu \in I_n} q_{\mu,N_\mu}(k), \tag{35}$$

$$q_{m,0}(k) = \beta_n^m(k) Q_n(k), \quad \forall m \in O_n, \tag{36}$$

where I_n is the set of links entering node n, O_n is the set of links leaving n, $Q_n(k)$ is the total traffic volume entering n at period k, $q_{m,0}(k)$ is the traffic volume that leaves n via outlink m, and $\beta_n^m(k)$ is the portion of $Q_n(k)$ that leaves the node through link m. $\beta_n^m(k)$ are the turning rates of node n and are assumed to be known for the entire time horizon. Equations (35) and (36) provide $q_{m,0}(k)$ required in (26) for $i = 1$.

The upstream influence of density and the downstream influence of speed at network nodes are taken under consideration by appropriate static models that provide the required terms in (26) and (27) for $i = 1$ and $i = N_m$ (Messmer and Papageorgiou, 1990).

5.3.2 *The constrained optimal control problem*

The coordinated ramp metering control problem is formulated as a dynamic optimal control problem with constrained control variables which can be solved numerically over a given time horizon. The general discrete-time formulation of the optimal control problem reads:

$$\text{minimize} \quad J = \vartheta[K] + \sum_{k=0}^{K-1} \varphi\big[\mathbf{x}(k), \mathbf{u}(k), \mathbf{d}(k)\big] \tag{37}$$

subject to

$$\mathbf{x}(k+1) = \mathbf{f}\big[\mathbf{x}(k), \mathbf{u}(k), \mathbf{d}(k)\big], \quad \mathbf{x}(0) = \mathbf{x}_0, \tag{38}$$

$$u_{i,\min} \leqslant u_i(k) \leqslant u_{i,\max}, \quad \forall i = 1, \ldots, m, \tag{39}$$

where K is the considered time horizon, $\mathbf{x} \in \Re^n$ is the state vector, $\mathbf{u} \in \Re^m$ is the vector of control variables, $\mathbf{d}$ is the vector of disturbances acting on the traffic process, and ϑ, φ are arbitrary, twice differentiable, nonlinear cost functions.

Based on the previous section, it may be seen that by substituting (27), (35), and (36) into (26); (29) into (27); (31)–(34) into (30), the traffic flow model equations take the form of Equation (38). In this case the state vector $\mathbf{x}$ consists of the densities $\rho_{m,i}$, the mean speeds $v_{m,i}$ of every segment i of every link m, and the queues w_o for every origin o. The control vector $\mathbf{u}$ consists of the ramp metering rates p_o of every on-ramp o under control, with $p_{o,\min} \leqslant p_o(k) \leqslant 1.0$ according to (39). Finally, the disturbance vector consists of all demands at each origin of the network and all turning rates at the network's bifurcations.

The chosen cost criterion aims at minimizing the Total Time Spent (TTS) of all vehicles in the network (including the waiting time experienced in the network queues). The cost criterion is as follows.

$$J = T \sum_k \left\{ \sum_m \sum_i \rho_{m,i}(k) L_m \lambda_m + \sum_o w_o(k) + a_f \sum_o \left[p_o(k) - p_o(k-1) \right]^2 + a_w \sum_o \psi\left[w_o(k) \right]^2 \right\} \tag{40}$$

with

$$\psi\left[w_o(k) \right] = \max\left\{ 0, w_o(k) - w_{o,\max} \right\}, \tag{41}$$

where the first two term in (40) account for the TTS while a_f, a_w are weighting factors. The term with weight a_f is included in the cost criterion to suppress high-frequency oscillations of the control trajectories. The last additional term is a penalty term included in the cost criterion in order to enable the control strategy to limit the queue lengths at the origins if and to the level desired. The parameters $w_{o,\max}$ are pre-determined constants and express the maximum permissible number of vehicles in origin o's queue.

A powerful numerical solution algorithm is used to solve this constrained discrete-time optimal control problem, see Papageorgiou and Marinaki (1995).

5.3.3 Application results

The previously described approach to network-wide optimal ramp metering has been applied to the Amsterdam ring-road with the use of AMOC.

The Amsterdam Orbital Motorway (A10) is shown in Figure 8. The A10 simultaneously serves local, regional, and inter-regional traffic and acts as a hub for traffic entering and exiting North Holland. There are four main connections with other motorways, the A8 at the North, the A4 at the South-West, the A2 at the South, and the A1 at the South-East. The A10 contains two tunnels, the Coen Tunnel at the North-West and the Zeeburg Tunnel at the East.

For the purposes of our study only the counter-clockwise direction of the A10, which is about 32 km long, is considered. There are 21 on-ramps on

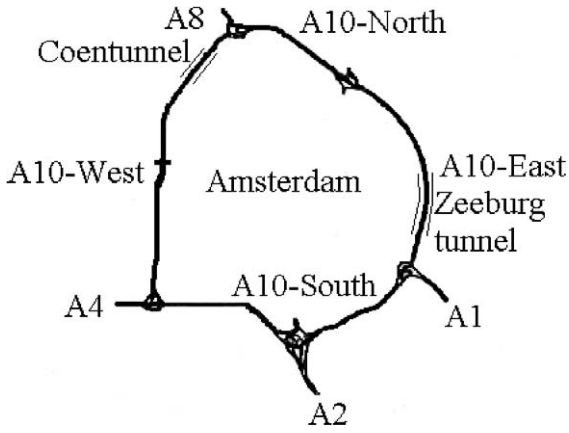


Fig. 8. The Amsterdam ring-road.

this motorway, including the connections with the A8, A4, A2, and A1 motorways, and a total number of 20 off-ramps, including the junctions with A4, A2, A1, and A8. It is assumed that ramp metering may be performed at each on-ramp, whereby the maximum permissible queue length for the on-ramps is set to 20 vehicles, while storage of 100 vehicles is permitted on each of the motorway-to-motorway ramps of A8, A4, A2, and A1.

The model parameters for this network were determined from validation of the network traffic flow model against real data taken from the motorways (Kotsialos et al., 1998).

The ring-road was divided in 76 segments with average length 421 m. This means that the state vector is 173-dimensional (including the 21 on-ramp queues). Since ramp metering is applied to all on-ramps, the control vector is 21-dimensional, while the disturbance vector is 43-dimensional. With a time step $T = 10$ s we have, for a horizon of 4 h, $K = 1440$ which results in a large-scale optimization problem with 279,360 variables.

5.3.4 The no-control case

The ring-road was studied for a time horizon of 4 hours, from 16:00 until 20:00, using realistic historical demands from the site. This time period includes the evening peak hour. In absence of any control measures, the ring-road is subject to recurrent congestion that is formed downstream of the junctions of A10 with A2 and A1 in A10-South. This congestion propagates backwards causing severe traffic delays in the A10-West. Figure 9(a) depicts the density propagation along the motorway segments (segment 0 is the first segment of A10-West after the junction of A10 with A8). The formation of large queues at the on-ramps can be seen in Figure 9(b) (on-ramp 0 corresponds to A8). As a result, the total time spent over the 4-h-horizon is equal to 13,226 veh h.

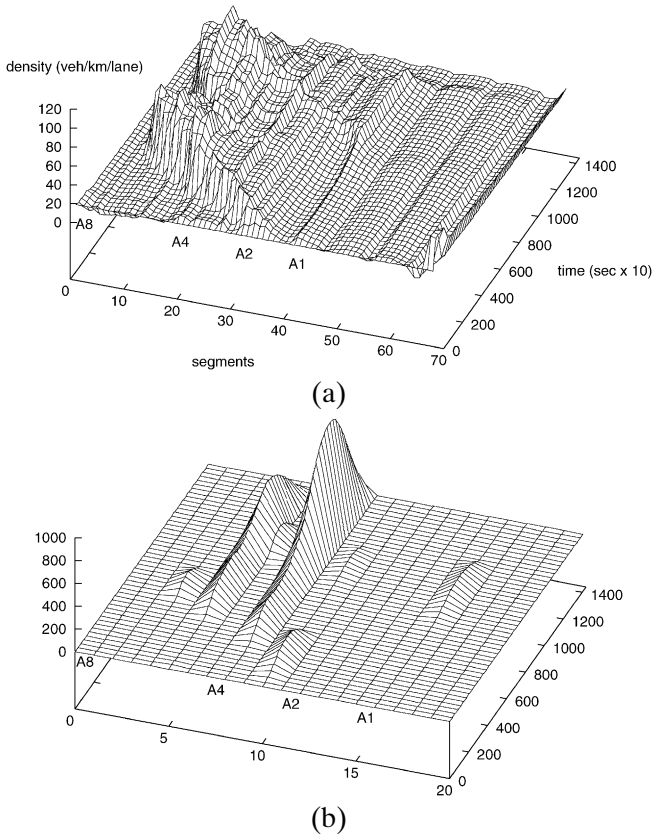


Fig. 9. No control: (a) Density, (b) on-ramp queues.

5.3.5 The control case

When ramp metering is performed at all on-ramps, the congestion is virtually lifted from the network (Figure 10(a)). The control strategy succeeds in establishing optimal uncongested traffic conditions on the A10-South and A10-West by applying ramp metering mainly at A1 and A2 at an early stage. In Figure 10(b), the queues are mainly occurring at A2 and A1 because these ramps have larger maximum permissible queues (100 vehicles). The control trajectories are depicted in Figure 10(c). The resulting total time spent is 8833 veh h, which is a 33.2% improvement compared to the no-control case.

A further improvement to the total time spent could be reached with larger maximum permissible queues. Had there been no queue constraints at all, the density profile of Figure 10(a) would be completely flat. In fact, the control strategy performs a trade-off between the queue lengths and the existence of congestion inside the network. Stricter queue constraints result in more degraded traffic conditions inside the motorway due to accordingly reduced control maneuverability.

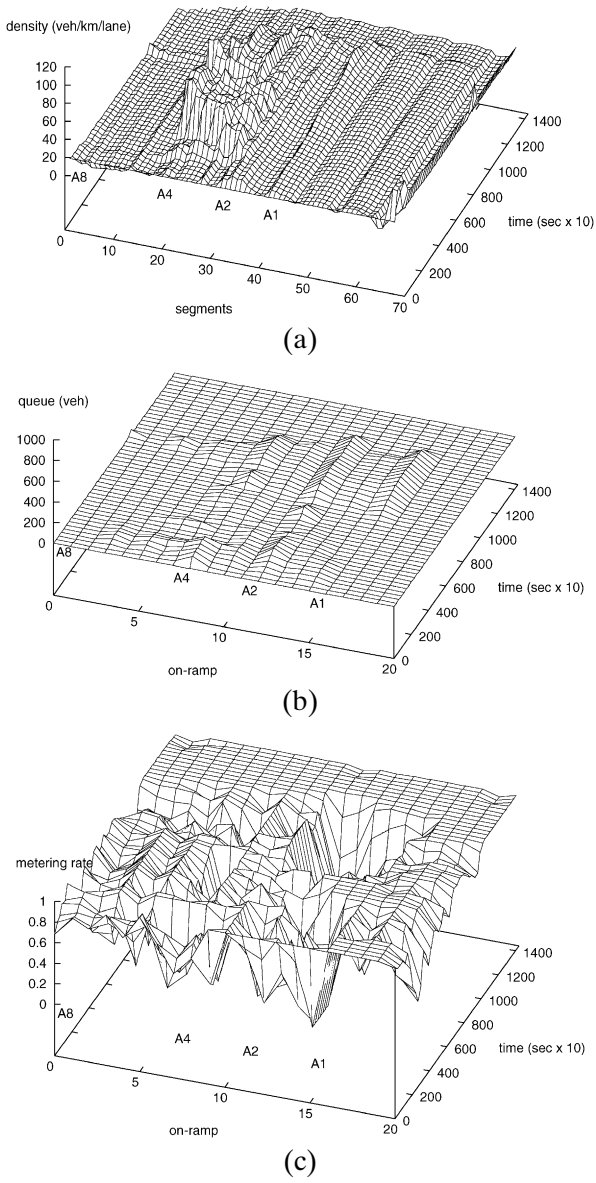


Fig. 10. Optimal control. (a) Density, (b) on-ramp queues, (c) optimal ramp metering rates.

The computation time required to obtain the optimal solutions is moderate and depends upon the search method used. The main part of the improvement is typically achieved very fast. The computation time for the 4-h-horizon is 20 min for the bulk of the 33.2% improvement (more than 30.6%) on a Sun Ultra5 with a Sparc IIi-360 MHz processor workstation.

5.4 Future directions

As in many other engineering disciplines, only a small portion of the significant methodological advancements in motorway network control have really been exploited in the field. It is beyond our scope to investigate and discuss the reasons behind this theory–practice gap, but administrative inertia, little competitive pressure in the public sector, the complexity of traffic control systems, limited realization of the improvement potential behind advanced methods by the responsible authorities, and limited understanding of practical problems by some researchers may have a role in this. Whatever the reasons, the major challenge in the coming decade is the deployment of advanced and efficient traffic control strategies in the field.

Regarding motorway networks, operational control systems of any kind are the exception rather than the rule. With regard to ramp metering, the main focus is frequently not on improving efficiency but on secondary objectives of different kinds. Most responsible traffic authorities and the decision makers are far from realizing the fact that advanced real-time ramp metering systems (employing optimal control algorithms) have the potential of changing dramatically the traffic conditions on today's heavily congested (hence strongly underutilized) motorways with spectacular improvements that may reach 50% reduction of the total time spent.

References

Ahmed, K.I., Ben-Akiva, M.E., Koutsopoulos, H.N., Mishalani, R.G. (1996). Models of freeway lane changing and gap-acceptance behaviour. In: Lesort, J.B. (Ed.), *Transportation and Traffic Theory: Proceedings 13th International Symposium of Transportation and Traffic Theory*. Pergamon/Elsevier, Lyon, France, pp. 505–515.

Al-Deek, H., Kanafani, A. (1993). Modeling the benefits of advanced traveler information systems in corridors with incidents. *Transportation Research C* 1 (4), 303–324.

Ataslar, B., Iftar, A. (1998). A decentralized control approach for transportation networks. In: *Preprints of the 8th IFAC Symposium on Large Scale Systems*, vol. 2. Patra, Greece, pp. 348–353.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A., Sugiyama, Y. (1995). Dynamical model of traffic congestion and numerical simulation. *Physical Review E* 51, 1035–1042.

Bell, M.C., Bretherton, R.D. (1986). Ageing of fixed-time traffic signal plans. In: *Proceedings of the 2nd International Conference on Road Traffic Control*. Institution of Electrical Engineers, London, UK.

Bellemans, T. (2003). Traffic control on motorways. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, May.

Ben-Akiva, M., De Palma, A., Kaysi, I. (1991). Dynamic network models and driver information systems. *Transportation Research A* 25 (5), 251–266.

Ben-Akiva, M., De Palma, A., Kaysi, I. (1996). The impact of predictive information on guidance efficiency: An analytical approach. In: Bianco, L., Toth, I. (Eds.), *Advanced Methods in Transportation Analysis*. Springer-Verlag, New York, pp. 413–432.

Benmohamed, L., Meerkov, S.M. (1994). Feedback control of highway congestion by a fair on-ramp metering. In: *Proceedings of the 33rd IEEE Conference on Decision and Control*, vol. 3. Lake Buena Vista, Florida, pp. 2437–2442.

Blinkin, M. (1976). Problem of optimal control of traffic flow on highways. *Automation and Remote Control* 37, 662–667.

Bode, K.R., Haller, W. (1983). Geschwindigkeitssteuerung auf der A7 zwischen den Autobahndreiecken Hannover-Nord und Walsrode. *Strassenverkehrstechnik* 27, 145–151 (in German).

Bolelli, A., Mauro, V., Perono, E. (1991). Models and strategies for dynamic route guidance, Part B: A decentralized, fully dynamic, infrastructure supported route guidance. In: *Proceedings of the DRIVE Conference in Advanced Telematics in Road Transports*, Brussels, Belgium, pp. 99–105.

Bonsall, P., Palmer, I. (1999). Route choice in response to variable message signs: Factors affecting compliance. In: Emmerink, R., Nijkamp, P. (Eds.), *Behavioural and Network Impacts of Driver Information Systems*. Ashgate Publishing Company, pp. 181–214. Chapter 9.

Bottom, J. (2000). Consistent anticipatory route guidance. PhD thesis, Dept. of Civil and Environmental Engineering, Massachusets Institute of Technology, Cambridge, MA, USA.

Bovy, P.H.L., Hoogendoorn, S.P. (2000). Ill-predictability of road-traffic congestion. In: Bell, M.G.H., Cassir, C. (Eds.), *Reliability of Transport Networks*. Research Studies Press, pp. 43–54.

Bovy, P.H., van der Zijpp, N.J. (1999). The close connection between dynamic traffic management and driver information systems. In: Emmerink, R., Nijkamp, P. (Eds.), *Behavioural and Network Impacts of Driver Information Systems*. Ashgate Publishing Company, pp. 355–370.

Brackstone, M., McDonald, M. (1999). Car-following: A historical review. *Transportation Research F* 2, 181–196.

Bretherton, D., Bowen, G., Wood, K. (2003). Effective urban traffic management and control – recent developments in SCOOT. In: *Proceedings of the 82nd TRB Annual Meeting*, Washington, DC.

Bretherton, D., Bodger, M., Baber, N. (2004). SCOOT – the future. In: *Proceedings of the 83rd TRB Annual Meeting*. Washington, DC.

Bretherton, R., Wood, K., Bowen, G.T. (1998). SCOOT version 4. In: *Proceedings of 9th International Conference on Road Transport Information and Control*. Institution of Electrical Engineers, London, UK.

Busch, F. (1996). Traffic telematics in urban and regional environments. In: *Proceedings of the Intertraffic Conference*. Amsterdam, The Netherlands.

Catling, I. (1989). AUTOGUIDE – electronic route guidance in the UK. In: Perrin, J.-P. (Ed.), *Control, Computers and Communications in Transportation*. Pergamon Press, Paris, France, pp. 269–276. Selected Papers from the IFAC/IFIP/IFORS Symposium.

Chandler, R.E., Herman, R., Montroll, E.W. (1958). Traffic dynamics: Studies in car following. *Operations Research* 6, 165–184.

Charbonnier, C., Farges, J.-L., Henry, J.-J. (1991). Models and strategies for dynamic route guidance, Part C: Optimal control approach. In: *Proceedings of the DRIVE Conference in Advanced Telematics in Road Transport*, Brussels, Belgium, pp. 106–112.

Cremer, M., Papageorgiou, M. (1981). Parameter identification for a traffic flow model. *Automatica* 17, 837–843.

Daganzo, C.F. (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research B* 28 (4), 269–287.

Daganzo, C.F. (1995). Requiem for second-order fluid approximations of traffic flow. *Transportation Research B* 29, 277–286.

Daganzo, C.F. (1997). A continuum theory of traffic dynamics for freeways with special lanes. *Transportation Research B* 31 (2), 83–102.

Daganzo, C.F. (2002a). A behavioural theory of multi-lane traffic flow. Part I: Long homogeneous freeway sections. *Transportation Research B* 36 (2), 131–158.

Daganzo, C.F. (2002b). A behavioural theory of multi-lane traffic flow. Part II: Merges and the on-set of congestion. *Transportation Research B* 36 (2), 159–169.

Department of the UK Environment, Transport and the Regions (DETR) (1999). The 'SCOOT' urban traffic control system. Traffic Advisory Leaflet 7/99, London, UK.

Diakaki, C., Papageorgiou, M. (1994). Design and simulation test of coordinated ramp metering control (METALINE) for A10-west in Amsterdam. Internal report 1994-2, Dynamic Systems and Simulation Laboratory, Technical University of Crete, Chania, Greece.

Diakaki, C., Papageorgiou, M., McLean, T. (2000). Integrated traffic-responsive urban corridor control strategy in Glasgow, Scotland – application and evaluation. *Transportation Research Record* 1727, 101–111.

Diakaki, C., Dinopoulou, V., Aboudolas, K., Papageorgiou, M., Ben-Shabat, E., Seider, E., Leibov, E. (2003). Extensions and new applications of the traffic signal control strategy TUC. *Transportation Research Record* 1856, 202–211.

Donati, F., Mauro, V., Roncoloni, G., Vallauri, M. (1984). A hierarchical decentralized traffic light control system-the first realisation: Progetto Torino. In: *Proceedings of the 9th World Congress of the International Federation of Automotive Control*, Budapest, Hungary, pp. 2853–2858.

Edie, L.C., Foote, R.S. (1958). Traffic flow in tunnels. *Highway Research Board Proceedings* 37, 334–344.

Edie, L.C., Foote, R.S. (1960). Effect of shock waves on tunnel traffic flow. *Highway Research Board Proceedings* 39, 492–505.

Emmerink, R., Axhausen, K.W., Nijkamp, P., Rietveld, P. (1995). The potential of information provision in a simulated road transport network with non-recurrent congestion. *Transportation Research C* 3 (5), 293–309.

Emmerink, R., Nijkamp, P., Rietveld, P., Van Ommeren, J.N. (1996). Variable message signs and radio traffic information: An integrated empirical analysis of drivers' route choice behavior. *Transportation Research A* 30 (2), 135–153.

Engelson, L. (1997). Self-fulfilling and recursive forecasts – an analytical perspective for driver information systems. In: *Preprints from the 8th Meeting of the International Association for Travel Behavior Research (IATBR '97)*. Workshop on Dynamics and ITS Response, Austin, TX, USA.

Esser, J., Neubert, L., Wahle, J., Schreckenberg, M. (1999). Microscopic online simulations of urban traffic. In: Ceder, A. (Ed.), *Transportation and Traffic Theory*. Proceedings of the 14th International Symposium of Transportation and Traffic Theory, Jerusalem, Israel. Pergamon/Elsevier, pp. 517–534.

Farges, J.L., Khoudour, K., Lesort, J.B. (1990). PRODYN: On site evaluation. In: *3rd International Conference on Road Traffic Control*. Institute of Electrical Engineers, London, UK, pp. 62–66.

Ferrari, P. (1989). The effect of driver behaviour on motorway reliability. *Transportation Research B* 23 (2), 139–150.

Forbes, T.W., Zagorski, H.J., Holshouser, E.L., Deterline, W.A. (1958). Measurement of driver reactions to tunnel conditions. *Highway Research Board Proceedings* 37, 345–357.

Gartner, N.N. (1991). Road traffic control: Demand responsive. In: Papageorgiou, M. (Ed.), *Concise Encyclopedia of Traffic and Transportation Systems*. Pergamon Press, pp. 386–391.

Hall, R.W. (1996). Route choice and advanced traveler information systems on a capacitated and dynamic network. *Transportation Research C* 4 (5), 289–306.

Head, K.L., Mirchandani, P.B., Sheppard, D. (1992). Hierarchical framework for real time traffic control. *Transportation Research Record* 1360, 82–88.

Hegyi, A. (2004). *Model Predictive Control for Integrating Traffic Control Measures*. PhD thesis, Delft University of Technology, Delft, The Netherlands.

Hegyi, A., De Schutter, B., Hellendoorn, J. (2003). MPC-based optimal coordination of variable speed limits to suppress shock waves in freeway traffic. In: *Proceedings of the 2003 American Control Conference*. Denver, CO, USA, pp. 4083–4088.

Helbing, D. (1997). *Verkehrsdynamik – neue physikalische Modellierungskonzepte*. Springer-Verlag (in German).

Helbing, D., Hennecke, A., Shvetsov, V., Treiber, M. (2001). MASTER: Macroscopic traffic simulation based on gas-kinetic, non-local traffic model. *Transportation Research B* 35 (2), 183–211.

Helly, W. (1959). Simulation of bottlenecks in single lane traffic flow. In: *Proceedings Symposium on Theory of Traffic Flow*, pp. 207–238.

Henn, V. (1995). Utilisation de la logique floue pour la modelisation microscopique du trafic routier. *La Revue du Logique Floue* (in French).

Herman, R., Montroll, E.W., Potts, R., Rothery, R.W. (1959). Traffic dynamics: Analysis of stability in car-following. *Operations Research* 1 (7), 86–106.

Hockney, R.W., Eastwood, J.W. (1988). *Computer Simulations Using Particles*. Hilger, Bristol, NY.

Hoffman, G. (1991). Up-to-the-minute information as we drive – how it can help road users and traffic management. *Transport Reviews* 11 (1), 41–61.

Holroyd, J., Hillier, J.A. (1979). The Glasgow experiment: PLIDENT and after. Technical Report LR384, Transport and Road Research Laboratory, Crowthorne, UK.

Hoogendoorn, S.P., Bovy, P.H.L. (1998). Optimal routing control using variable message signs. In: Bovy, P.H.L. (Ed.), *Motorway Flow Analysis: New Methodologies and Recent Empirical Findings*. Delft Univ. Press, pp. 237–263.

Hoogendoorn, S.P., Bovy, P.H.L. (1999). Multiclass macroscopic traffic flow modelling: A multilane generalisation using gas-kinetic theory. In: Ceder, A. (Ed.), *Transportation and Traffic Theory*. Proceedings of the 14th International Symposium of Transportation and Traffic Theory, Jerusalem, Israel. Pergamon/Elsevier, pp. 27–50.

Hoogendoorn, S.P., Bovy, P.H.L., van Lint, J.W.C. (2002). Short-term prediction of traffic flow conditions in a multilane multiclass network. In: Taylor, M.A.P. (Ed.), *Transportation and Traffic Theory in the 21st Century: Proceedings of the 15th International Symposium on Transportation and Traffic Theory*. Pergamon/Elsevier, Oxford, pp. 625–651.

Hunt, P.B., Robertson, R.D., Bretherton, R.D., Winton, R.I. (1981). SCOOT – a traffic responsive method of co-ordinating signals. Technical Report 1014, Transport and Road Research Laboratory.

Jeffery, D.J. (1981). The potential benefits of route guidance. Technical Report LR997, Transport and Road Research Laboratory.

Jhaveri, C.S., Perrin, J., Martin, P.T. (2003). SCOOT adaptive signal control: An evaluation of its effectiveness over a range of congestion intensities. In: *Proceedings of the 82nd TRB Annual Meeting*. Washington, DC.

Kantowitz, B.H., Hanowski, R.J., Kantowitz, S.C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors* 39 (2), 164–176.

Kaufman, D., Smith, R., Wunderlich, K. (1998). User-equilibrium properties of fixed points in iterative dynamic routing/assignment methods. *Transportation Research C* 6 (1–2), 1–16.

Kaysi, I., Ben-Akiva, M., De Palma, A. (1995). Design aspects of advanced traveler information systems. In: Gartner, N., Improta, G. (Eds.), *Urban Traffic Networks: Dynamic Flow Modeling and Control*. Springer-Verlag, pp. 59–81.

Kerner, B.S. (1999). Theory of congested traffic flow: Self-organization without bottlenecks. In: Ceder, A. (Ed.), *Transportation and Traffic Theory*. Proceedings of the 14th International Symposium of Transportation and Traffic Theory, Jerusalem, Israel. Pergamon/Elsevier, pp. 147–172.

Kerner, B.S., Rehborn, H. (1997). Experimental properties of phase-transitions in traffic flow. *Physical Review Letters* 79 (20), 4030–4033.

Kerner, B.S., Aleksic, M., Rehborn, H. (2000). Automatic tracing and forecasting of moving traffic jams using predictable features of congested traffic flow. In: Schnieder, E., Becker, U. (Eds.), *Control in Transportation Systems 2000: Proceedings of the 9th IFAC Symposium on Control in Transportation Systems*. Braunschweig, Germany, pp. 501–506.

Khattak, A., Polydoropoulou, A., Ben-Akiva, M. (1996). Modeling revealed and stated pre-trip travel response to ATIS. *Transportation Research Record* 1537, 46–54.

Khattak, A.J., Yim, Y., Stalker, L. (1999). Does travel information influence commuter and noncommuter behavior? Results from the San Francisco Bay Area TravInfo project. *Transportation Research Record* 1694, 48–58.

Kikuchi, C., Chakroborty, P. (1992). Car following model based on a fuzzy inference system. *Transportation Research Record* 1365 (1992), 82–91.

Klar, A., Wegener, R. (1999a). A hierarchy of models for multilane vehicular traffic I: Modelling. *SIAM Journal of Applied Mathematics* 59 (3), 983–1001.

Klar, A., Wegener, R. (1999b). A hierarchy of models for multilane vehicular traffic II: Numerical investigations. *SIAM Journal of Applied Mathematics* 59 (3), 1002–1011.

Kotsialos, A., Pavlis, Y., Middelham, F., Diakaki, C., Vardaka, G., Papageorgiou, M. (1998). Modelling of the large scale motorway network around amsterdam. In: *Preprints of the 8th IFAC Symposium on Large Scale Systems*, vol. 2. Patra, Greece, pp. 354–360.

Kotsialos, A., Papageorgiou, M., Messmer, A. (1999). Optimal coordinated and integrated motorway network traffic control. In: Ceder, A. (Ed.), *Transportation and Traffic Theory*. Proceedings of the 14th International Symposium of Transportation and Traffic Theory (ISTTT), Jerusalem, Israel. Pergamon/Elsevier, pp. 621–644.

Kotsialos, A., Papageorgiou, M., Mangeas, M., Haj-Salem, H. (2002). Coordinated and integrated control of motorway networks via nonlinear optimal control. *Transportation Research C* 10 (1), 65–84.

Koutsopoulos, H.N., Lotan, T. (1989). Effectiveness of motorist information systems in reducing traffic congestion. In: *Proceedings of the Conference on Vehicle Navigation and Information Systems (VNIS)*, Toronto, Canada, pp. 275–281.

Kraan, M., Mahmassani, H.S., Huynh, N. (2000). Interactive survey approach to study traveler responses to ATIS for shopping trips. In: *Proceedings of the 79th Transportation Research Board Annual Meeting*. Washington, DC, USA.

Kühne, R.D. (1991). Traffic patterns in unstable traffic flow on freeways. In: Brannolte, U. (Ed.), *Highway Capacity and Level of Service*. Proceedings of the International Symposium on Highway Capacity, Karlsruhe, 24–27 July 1991. Balkema, Rotterdam, pp. 211–223.

Lebacque, J.P. (2002). A two-phase extension of the LWR model based on the boundness of traffic acceleration. In: Taylor, M.A.P. (Ed.), *Transportation and Traffic Theory in the 21St Century*. Proceedings of the 15th International Symposium on Transportation and Traffic Theory, Adelaide, Australia. Pergamon/Elsevier, pp. 697–718.

Leutzbach, W. (1988). *An Introduction to the Theory of Traffic Flow*. Springer-Verlag, Berlin, Germany.

Leutzbach, W. (1991). Measurements of mean speed time series autobahn A5 near Karlsruhe, Germany. Technical report, Institute of Transport Studies, University of Karlsruhe, Germany.

Leutzbach, W., Wiedemann, R. (1986). Development and applications of traffic simulation models at the Karlsruhe Institute fuer Verkehrswesen. *Traffic Engineering and Control* 27 (5), 270–278.

Lighthill, M.J., Whitham, G.B. (1955). On kinematic waves II: A traffic flow theory on long crowded roads. *Proceedings of the Royal Society of London Series A* 229, 317–345.

Lindley, J. (1987). Urban freeway congestion: Quantification of the problem and effectiveness of potential solutions. *ITE Journal* 57(1).

Logghe, S. (2003). Dynamic modelling of heterogeneous vehicular traffic. PhD thesis, Catholic University of Leuven, Belgium.

Lowrie, P.R. (1982). The Sydney coordinated adaptive traffic system; principles, methodology, algorithms. In: *Proceedings of Institute of Electrical Engineers Internacional Conference on Road Traffic Signalling* London, pp. 67–70.

Ludmann, J. (1998). Beeinflussung des Verkehrsablaufs auf Strassen – Analyse mit dem fahrzeugorientierten Verkehrssimulationsprogramm PELOPS. Technical report, Schriftenreihe Automobiltechnik, Institut für Kraftfahrwesen, Aachen, Germany (in German).

Mahmassani, H., Jayakrishnan, R. (1991). System performance and user response under real-time information in a congested traffic corridor. *Transportation Research A* 25 (5), 293–307.

Mahmassani, H., Liu, Y.H. (1999). Dynamics of commuting decision behavior under advanced traveler information systems. *Transportation Research C* 7 (2–3), 91–107.

Mannering, F.L., Kim, S.-G., Barfield, W., Ng, L. (1994). Statistical analysis of commuters' route, mode and departure time flexibility. *Transportation Research C* 2 (1), 35–47.

Masher, D.P., Ross, D.W., Wong, P.J., Tuan, P.L., Zeidler, A., Peracek, S. (1975). Guidelines for design and operating of ramp control systems. Technical Report NCHRP 3-22, SRI Project 3340, Standford Research Institute, SRI, Menid Park, California, USA.

May, A.D. (1990). *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, NJ.

McDonald, M., Hounsell, N.B. (1991). Road traffic control: TRANSYT and SCOOT. In: Papageorgiou, M. (Ed.), *Concise Encyclopedia of Traffic and Transportation Systems*. Pergamon, pp. 400–408.

Messmer, A., Papageorgiou, M. (1990). METANET: A macroscopic simulation program for motorway networks. *Traffic Engineering and Control* 31 (8), 466–470; erratum 31 (9), 549.

Michaels, R.M. (1963). Perceptual factors in car following. In: *Proceedings of the 2nd Symposium on Theory of Road Traffic Flow*. Paris, France, pp. 44–59.

Minderhoud, M.M. (1999). Supported Driving: impacts on motorway traffic Flow. PhD thesis, Delft University of Technology, Delft, The Netherlands.

Mirchandani, P., Head, L., Knyazyan, A., Wu, W. (2001). An approach towards the integration of bus priority and traffic adaptive signal control. In: *Proceedings of the 80th TRB Annual Meeting*. Washington, DC.

Moreno-Banos, J.C., Papageorgiou, M., Schaffner, C. (1993). Integrated optimal flow control in traffic networks. *European Journal of Operations Research* 71, 317–323.

Munjal, P., Pahl, J. (1969). An analysis of the Boltzmann-type statistical models for multi-lane traffic flow. *Transportation Research* 3, 151–163.
Nagel, K. (1996). Particle hopping models and traffic flow theory. *Physical Review E* 53 (5), 4655–4672.
Nagel, K. (1998). From particle hopping models to traffic flow theory. *Transportation Research Record* 1644, 1–9.
Nelson, P., Sopasakis, A. (1998). The Prigogine–Herman kinetic model predicts widely scattered traffic flow data at high concentrations. *Transportation Research B* 32 (8), 589–604.
Nelson, P., Bui, D.D., Sopasakis, A. (1997). A novel traffic stream model deriving from a bimodal kinetic equilibrium. In: Papageorgiou, M., Pouliezos, A. (Eds.), *Transportation Systems 1997*. Proceedings of the 8th IFAC Symposium on Transportation Systems, Chania, Crete, Greece. Pergamon, pp. 799–804.
Newell, G.F. (1961). A theory of traffic flow in tunnels. In: Herman, R. (Ed.), *Theory of Traffic Flow*, pp. 193–206.
Ozbay, K., Datta, A., Kachroo, P. (2001). Modeling route choice behavior using stochastic learning automata. In: *Proceedings of the 80th Transportation Research Board Annual Meeting*. Washington, DC, USA.
Papageorgiou, M. (1980). A new approach to time-of-day control based on a dynamic freeway traffic model. *Transportation Research B* 14, 349–360.
Papageorgiou, M. (1990). Dynamic modeling, assignment, and route guidance in traffic networks. *Transportation Research B* 24, 471–495.
Papageorgiou, M., Marinaki, M. (1995). A feasible direction algorithm for the numerical solution of optimal control problems. Internal report 1995-4, Dynamic Systems and Simulation Laboratory, Technical University of Crete, Chania, Greece.
Papageorgiou, M., Blosseville, J.M., Hadj-Salem, H. (1990). Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris Part II: Coordinated on-ramp metering. *Transportation Research A* 24, 361–370.
Papageorgiou, M., Haj-Salem, H., Blossville, J.M. (1991). ALINEA: A local feedback control law for on-ramp metering. *Transportation Research Record* 1320, 58–64.
Papageorgiou, M., Haj-Salem, H., Middelham, F. (1998). ALINEA: A local ramp metering: Summary of field results. *Transportation Research Record* 1603, 90–98.
Paveri-Fontana, S.L. (1975). On Boltzmann-like treatments for traffic flow: A critical review of the basic model and an alternative proposal for dilute traffic analysis. *Transportation Research B* 9, 225–235.
Pavlis, Y., Papageorgiou, M. (1999). Simple decentralized feedback strategies for route guidance in traffic networks. *Transportation Science* 33 (3), 264–278.
Payne, H.J. (1971). Models for freeway traffic and control. In: G.A. Bekey, (Ed.), *Mathematical Models of Public Systems 1*, pp. 51–61.
Peirce, S., Lappin, J. (2002). Evolving awareness, use, and opinions of Seattle region commuters concerning traveler information: Findings from the puget sound transportation panel survey, 1997 and 2000. In: *Proceedings of the 82nd Transportation Research Board Annual Meeting*.
Pignataro, L.J. (1973). *Traffic Engineering – Theory and Practice*. Prentice-Hall, Englewood Cliffs, NJ.
Pipes, L.A. (1953). An operational analysis of traffic dynamics. *Journal of Applied Physics* 24 (1), 274–287.
Polydoropoulou, A., Ben-Akiva, M. (1999). The effect of Advanced Traveler Information Systems (ATIS) on travelers behavior. In: Emmerink, R., Nijkamp, P. (Eds.), *Behavioral and Network Impacts of Driver Information Systems*. Wiley.
Polydoropoulou, A., Ben-Akiva, M., Khattak, A., Lauprete, G. (1996). Modeling revealed and stated en-route travel response to ATIS. *Transportation Research Record* 1537, 38–45.
Pozybill, M. (1998). Ist Verkehr chaotisch? *Strassenverkehrstechnik* 10, 538–545 (in German).
Prigogine, I., Herman, R. (1971). *Kinetic Theory of Vehicular Traffic*. Elsevier, New York.
Rekersbrink, A. (1995). Mikroskopische Verkehrssimulation mit Hilfe der Fuzzy Logic. *Strassenverkehrstechnik* 2, 68–74 (in German).
Rilett, L.R., van Aerde, M., MacKinnon, G., Krage, M. (1991) Simulating the TravTek route guidance logic using the INTEGRATION traffic model. In: *Proceedings of the Conference on Vehicle Navigation and Information Systems. SAE Conference Proceedings No. 253*, Part 2. SAE, Warrendale, PA, USA, pp. 775–787.

Robertson, D.I. (1997). The TRANSYT method of coordinating traffic signals. *Traffic Engineering and Control* 38 (2), 76–77.

Schwartz, S.C., Tan, H.H. (1977). Integrated control of freeway entrance ramps by threshold regulation. In: *Proceedings IEEE Conference of Decision and Control*, pp. 984–986.

Smoulders, S. (1990). Control of freeway traffic flow by variable message signs. *Transportation Research B* 24, 111–132.

Tampére, C., van Arem, B., Hoogendoorn, S.P. (2002). Gas kinetic traffic flow modelling including continuous driver behaviour models. In: Bovy, P.H.L. (Ed.), *Proceedings of the 7th Annual TRAIL Conference*. Delft Univ. Press, Delft, The Netherlands.

Treiterer, J., Myers, J.A. (1974). The hysteresis phenomena in traffic flow. In: Buckley, D.J. (Ed.), *Proceedings of the 6th International Symposium on Transportation and Traffic Flow Theory*. A.M. & A.W. Reed Private Limited, Artarmon, NSW, pp. 13–38.

Valdes, D., Paz, A. (2004). An integration of adaptive traffic control and travel information. In: *Proceedings of the 83rd TRB Annual Meeting*. Washington, DC.

van Aerde, M. (1994). INTEGRATION: A model for simulating integrated traffic networks. Technical report, Transportation Systems Research Group, Queens University, Canada.

van Arem, B., Hogema, J.H. (1995). The microscopic simulation model MIXIC 1.2. Technical Report 1995-17b, TNO-INRO, Delft, The Netherlands.

Verweij, H.D. (1985). Congestion warning and traffic flow operations. Technical report, Dutch Ministry of Traffic and Transportation, Delft, The Netherlands (in Dutch).

Watling, D., van Vuren, T. (1993). The modelling of dynamic route guidance systems. *Transportation Research C* 1 (2), 159–182.

Wattleworth, J.A. (1965). Peak-period analysis and control of a freeway system. *Highway Research Record* 157, 1–21.

Wu, N., Brilon, W. (1999). Cellular automata for highway traffic flow simulation. In: Ceder, A. (Ed.), *Transportation and Traffic Theory*. Proceedings of the 14th International Symposium of Transportation and Traffic Theory, Jerusalem, Israel. Pergamon/Elsevier, pp. 1–18 (abbreviated presentations).

Wunderlich, K.E., Hardy, M.H., Larkin, J.J., Shah, V.P. (2001). On-time reliability impacts of Advanced Traveler Information Services (ATIS): Washington, DC case study. Technical report, Mitretek Systems.

Yang, H. (1998). Multiple equilibrium behavior and advanced traveler information systems with endogenous market penetration. *Transportation Research B* 32 (3).

Yim, Y., Miller, M.A. (2000). Evaluation of the TravInfo field operational test. Technical report, Institute of Transportation Studies, University of California, Berkeley. California PATH Program.

Young, P.C., Taylor, J., Chotai, A., Whittaker, J. (1997). A non-minimal state variable feedback approach to co-ordinated ramp metering. In: Kotsialos, A. (Ed.), *DACCORD Deliverable D 6.1 - Cordinated Control*, vol. 6.1. European Commission, Brussels, Belgium.

Yuan, L.S., Kreer, J.B. (1968). An optimal control algorithm for ramp metering of urban freeways. In: *Proceedings of the 6th IEEE Annual Allerton Conference on Circuit and System Theory*. Allerton, Illinois, USA.

Zackor, H. (1972). Beurteilung verkehrsabhangiger Geschwindigkeitsbeschrankungen auf Autobahnen. *Forschung Strassenbau und Strassenverkehrstechnik* 128, 1–61.

Zackor, H., Balz, W. (1984). Verkehrstechnische Funktionen des Leitsystems. *Strassenverkehrstechnik* 28, 5–9.

Zhang, H.M. (2003). Driver memory, traffic viscosity and a viscous vehicular traffic flow model. *Transportation Research B* 37 (1), 27–41.

Subject Index

D

E

S

T