Risk and Decision Analysis in Maintenance Optimization and Flood Management

Edited by M.J. Kallen & S.P. Kuniewski

Papers presented at the symposium
in remembrance of prof. Jan M. van Noortwijk
on November 24, 2009 in Delft, the Netherlands

Risk and Decision Analysis
in
Maintenance Optimization and Flood Management

This page intentionally left blank

Risk and Decision Analysis in Maintenance Optimization and Flood Management

Edited by M.J. Kallen and S.P. Kuniewski

IOS Press

© 2009 The authors and IOS Press. All rights reserved.

ISBN 978-1-60750-068-1

Published by IOS Press under the imprint Delft University Press

Publisher
IOS Press BV
Nieuwe Hemweg 6b
1013 BG Amsterdam
The Netherlands
tel.: +31-20-688 3355
fax: +31-20-687 0019
www.iospress.nl
www.dupress.nl

Cover design by Maarten-Jan Kallen and created by Kuniewska Sylwia. Artwork on front cover by Jan van Dijk.

LEGAL NOTICE
The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

This book is dedicated to the memory of

Jan Maarten van Noortwijk (⋆1961-†2008)

This page intentionally left blank

Table of Contents

This page intentionally left blank

Foreword

In his position as professor at the faculty of Electrical Engineering, Mathematics and Computer Science at the Delft University of Technology, Jan van Noortwijk had a simple goal: to apply mathematical modeling techniques to problems in civil engineering. In particular, he aimed to make advanced decision-theoretic models accessible to engineers in other fields such as civil and mechanical engineering. Most of his work involved the application of probability theory to problems in maintenance optimization and the management of risks due to flooding. The inherent uncertainty involved with the current and future state of structures and systems requires a sound methodology for quantifying these uncertainties.

This book presents some of the latest developments in these areas by leading researchers at academic institutions and practitioners in various lines of work. The contributions will be presented during a one-day symposium on November 24, 2009 in Delft, the Netherlands. Both this book and the symposium are a tribute to the legacy of professor Jan van Noortwijk.

First and foremost we are indebted to the authors for their enthousiastic response to the call for papers and the significant effort they have put into finishing their contributions within a very short period of time. We extend our appreciation to the scientific committee, being Tim Bedford, Christophe Bérenguer, Rommert Dekker, Pieter van Gelder, Antoine Grall, Matthijs Kok, Tom Mazzuchi, Robin Nicolai, Martin Newby, and Hans van der Weide for their swift reviews. We would also like to thank Ton Botterhuis and Karolina Wojciechowska for additional reviewing and editing of a number of contributions.

At the time of writing, the symposium has been made possible by the organizing institutions, HKV Consultants and the Delft University of Technology, as well as the Nederlandse Vereniging voor Risicoanalyse en Bedrijfszekerheid (NVRB), Universiteitsfonds Delft, the Netherlands Organization for Applied Scientific Research (TNO), and the organizing committee of the 7th International Probabilistic Workshop (November 25-26, 2009 in Delft).

The editors,
Maarten-Jan Kallen and Sebastian Kuniewski
Delft, September 23, 2009.

This page intentionally left blank

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

The work of professor Jan van Noortwijk (1961-2008): an overview

Maarten-Jan Kallen*and Matthijs Kok
– HKV Consultants, Lelystad, the Netherlands

Abstract. We give an overview of the research and publications by professor Jan van Noortwijk starting from his graduation at the Delft University of Technology in 1989 up to his death on September 16, 2008. The goal of this overview is to list all of his scientific publications and to put these in a historical perspective. We show how his Ph.D. thesis was a stepping stone to the two primary fields in which he did most of his later work: maintenance optimization and the management of risks due to flooding.

1 THE FORMATIVE YEARS: 1988 TO 1995

In 1988 Jan was an undergraduate student at the Delft University of Technology. At that time, he was majoring in applied mathematics at the faculty of Mathematics and Computer Science and working on his Master's thesis under the supervision of Roger Cooke. Rommert Dekker, now a professor at the Erasmus University in Rotterdam but at that time working in the department of Mathematics and Systems Engineering at the research laboratorium of Royal Dutch/Shell in Amsterdam, approached Roger Cooke with a problem they were having with a decision support system for maintenance optimization called PROMPT-II [1].

The PROMPT system was designed for optimal opportunity-based preventive maintenance. One problem was that the system used lifetime distributions requiring an amount of data which was unavailable at that time. Their attempts at elicitation of this data using expert opinion among their engineers resulted in many inconsistencies between estimates. During his internship at Royal Dutch/Shell, where he was supervised by Rommert Dekker and Thomas Mazzuchi, Jan van Noortwijk developed methods to elicit expert opinion on reliability data in a structured manner and to combine these estimates into a consensus distribution for the lifetime of a component. This research resulted in his Master's thesis [2] with which he graduated from the university in 1989. It also resulted in his first and most cited publication in a scientific journal [3]. Another student of Roger Cooke,

*corresponding author: HKV Consultants, P.O. Box 2120, 8203 AC Lelystad, the Netherlands; telephone: +31-(0)320 294 256, fax: +31-(0)320 253 901, e-mail: m.j.kallen@hkv.nl.

René van Dorp, continued Jan's work at Shell and implemented the elicitation procedure suggested by Jan. He also developed feedback for the elicitation procedure, which included feedback for evaluating the optimal maintenance interval given the elicited lifetime distribution [4].

Both Roger Cooke and Rommert Dekker suggested to Jan that he should pursue a doctoral degree at the university, but Jan went to work for the Dr. Neherlab in Leidschendam, which was the research laboratory of the Dutch national telecommunications company. During the short period that he worked there (up to August 1990), he co-authored one conference paper [5]. In September 1990, Jan returned to the university in Delft and became a graduate student, initially with professor Freek Lootsma in the Operations Research chair, but later with Roger Cooke whom became a professor in the Risk Analysis and Decision Theory chair. Around this time, Matthijs Kok at Delft Hydraulics (now Deltares), and a former graduate student of prof. Lootsma, was setting up a research program on the optimal maintenance of hydraulic structures. After a meeting with Roger Cooke and Jan van Noortwijk, Matthijs appointed Jan as a contractor. Jan held his position at the university until June 1995 and obtained his doctoral degree on the 28th of May in 1996 with his thesis *Optimal maintenance decisions for hydraulic structures under isotropic deterioration*; see [6] and Figure 1.

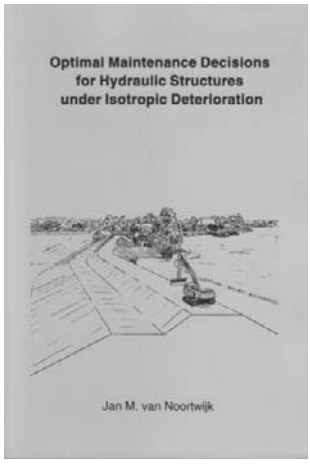


FIGURE 1. the front cover of Jan van Noortwijk's Ph.D. thesis also known as the 'little yellow book' due to the bright yellow color of the cover.

The contract work for Delft Hydraulics provided a unique opportunity for Jan to work on real life problems and almost every chapter from his thesis was later published in a scientific journal. The four primary applications discussed in his thesis are: optimal sand nourishment decisions for the Dutch coastline [7], optimal maintenance decisions for dykes [8], for berm

breakwaters [9], and for the sea-bed protection of the Eastern-Scheldt storm surge barrier [10]. The problem of optimally inspecting the block-mats of the Eastern-Scheldt barrier was also the topic of a chapter in a book published in conjunction with a workshop which was organized to celebrate his Ph.D. thesis; see [11]. These mats prevent possible instability of the piers in the barrier due to erosion and must be inspected periodically to check for the presence of scour holes. Jan proposed a Poisson process for the random occurrence of these scour holes and a gamma process for the stochastic expansion of the size of the holes once they have appeared.

From a theoretical point of view, his most important contribution by his the use of the gamma process to model uncertain deterioration over time. His motiviation for this was not only the fact that the increments of this particular stochastic process are non-negative, which makes the process of deterioration monotonically increasing, but also that it could be characterized by the only (subjective) information which is commonly available, namely the limiting average rate of deterioration. This feature makes the gamma process fit within the operational Bayesian approach advocated by Max Mendel and Richard Barlow; see [12] and [13]. The basic thought behind this approach is that any model should be designed such that prior information need only be given over parameters with an operational meaning. Jan visited Max and Dick as a visiting scholar at the University of California at Berkeley in 1992 and this ultimately gave direction to the mathematical aspects of his research [14]. These aspects are the topics of the second and third chapter in Jan's Ph.D. thesis.

In the second chapter of his Ph.D. thesis, Jan discusses a Bayesian isotropic failure model which is based on two assumptions: (1) the order in which the increments appear is irrelevant (i.e., they are exchangeable) and (2) given the average amount of deterioration per unit time, the decision maker is indifferent to the way this average is obtained (i.e., the amounts of deterioration are ℓ_1-isotropic, which implies exchangeability). The latter may also be stated as follows: all combinations leading to the same average have the same degree of belief for the decision-maker. This chapter was later published in the *European Journal of Operations Research* [15]. Note that the assumption of ℓ_1-isotropic deterioration implies that the expected amount of deterioration is linear in time. The third chapter in his Ph.D. thesis characterizes the general gamma process in terms of sufficiency and isotropy. This work was done together with Jolanta Misiewicz from the University of Zielona Gora in Poland, which was later published in the *Journal of Mathematical Sciences* [16]. The ninth and last chapter of his thesis contains the results of a follow-up on the research he did for his M.Sc. thesis and which was reported in his first journal publication [3]. This chapter, which was later published in the *Journal of Quality in Maintenance Engineering* [17], proposes the use of the Dirichlet distribution as a discrete lifetime distribution, which can be used when experts give estimates of lifetimes in the form of a histogram.

Jan van Noortwijk also wrote several reports for Delft Hydraulics. The first was an inventory of problems for his research [18]. In 1991, Leo Klatter from the Ministry of Transport, Public Works and Water Management asked Matthijs Kok to research how to optimally maintain parts of hydraulic structures which were located below the waterline [19]. The report included an analysis of the Eastern-Scheldt barrier, which would become one of the real life applications in Jan's thesis. In [20], Jan used the method of paired comparisons to rank, amongst other variables, the various designs of the bridge now known as the Erasmus bridge in Rotterdam. For each layout, ship pilots (i.e., the experts), were asked whether it would be easier or more difficult to navigate relative to the other layouts. However, his most important work for Delft Hydraulics would become the work he and Matthijs Kok did in 1994 for the Committee Flood Disaster Meuse, also known as "Boertien-II", which will be discussed in Section 5.

2 THE PROFESSIONAL CAREER: 1995 TO 2008

On September 1, 1995, Hans Hartong, Matthijs Kok and Kees Vermeer founded *HKV Lijn in water B.V.* (English: *HKV Consultants*) in the city of Lelystad in the Netherlands. One month later, Jan van Noortwijk joined them as their first employee. From this point on, his work would focus on roughly two areas: maintenance optimization of man-made structures and systems, and the assessment and management of risks related to natural hazards such as coastal and fluvial flooding. This is best expressed by the longstanding relationship with people at two divisions of the Directorate-General for Public Works and Water Management, namely the Centre for Public Works and the Centre for Water Management. A detailed account of his achievements in both subject areas is the topic of Sections 4 and 5.

On May 1, 2000, at which time the company had grown to 36 employees, Jan became the head of the newly formed *Risk and Safety* group. In the Netherlands, he had quickly gained recognition by his peers as being a leading expert in his field. Combined with the multitude of publications detailing his pioneering work in both his areas of interest, this led to his appointment as a part-time professor at the Delft University of Technology. There, he would join professor Roger Cooke at the faculty of Electrical Engineering, Mathematics and Computer Science as the head of the chair *Mathematical Aspects of Risk Analysis*. On the day of his death, September 16, 2008, HKV Consultants had grown to 62 people and the Risk and Safety group had grown from 8 to 16 members.

In the following sections, we describe the work of Jan van Noortwijk in three subject areas: uncertainty and sensitivity analysis, maintenance optimization, and flood risk management.

3 UNCERTAINTY AND SENSITIVITY ANALYSIS

Around the time that Jan was finishing his Ph.D. thesis and starting his work at HKV Consultants, he worked on general developments in the theory of uncertainty and sensitivity analysis. In particular, he was involved with the development of a software tool called *Uncertainty analysis with Correlations* (UNICORN) together with Roger Cooke at the Delft University of Technology. He co-authored several papers related to this tool together with Roger Cooke. In [21, 22] they discuss graphical methods for use in uncertainty and sensitivity analyses. One of these methods is the use of so-called *cobweb* plots for the visual display of correlated random variables. These were used in their uncertainty analysis of the reliability of dike-ring areas in the Netherlands [23].

4 MAINTENANCE OPTIMIZATION

Jan's work in deterioration modeling and maintenance optimization was largely influenced by his work for Leo Klatter and Jaap Bakker at the Centre for Public Works and by his position as professor at the university in Delft. The Centre for Public Works is essentially a knowledge centre for issues regarding the management of important civil infrastructures, such as the national roads, bridges, sluices, storm-surge barriers, etcetera. The two subjects that Jan was most involved with, were the management of road bridges and the maintenance of coating systems on steel structures. He would complete several projects for the Centre, but he was also hired as a contractor for a long period of time during which he spent about one day a week at the offices of the Centre in Utrecht.

4.1 Lifetime-extending maintenance model

Together with Jaap Bakker and Andreas Heutink at the Centre for Public Works and several colleagues at HKV, Jan developed the lifetime-extending maintenance (LEM) model [24] and the inspection-validation model [25, 26]. The LEM model originated from a spreadsheet module for the calculation of the net present value of future expenditures, which was made together with Harry van der Graaf. Given the available information on the rate of deterioration and the uncertainty in the expected lifetime of the object, the LEM model can be used to balance the costs of lifetime-extending maintenance versus complete replacements. It does so by comparing the life-cycle costs of two maintenance policies: one with periodic (imperfect) repairs which extend the lifetime of the object and one with only periodic replacements which bring the object back to an as-good-as-new state. The inspection-validation module can be used to update the deterioration process in the LEM model, which is based on the gamma process, with information gained by inspection of the state of the object.

4.2 Bridge management and life-cycle costing

Most of his work for the Centre concerned the topic of bridge management. Many of the bridges in the national road network of the Netherlands were built in the late 1960's and early 1970's. As many of these structures require significant maintenance and retrofitting after approximately 40 to 50 years, the Centre expects that a large number of bridges will have to be maintained in the near future and that this would put severe pressure on the already shrinking budget for the management of national infrastructures. The challenge is therefore to prioritize maintenance actions and to communicate the necessity of large-scale repairs to policy makers and to the public. Prioritization should be based on the principle of life-cycle costing (LCC), as current decisions affect future maintenance requirements. In order to apply this principle, it is necessary to have an estimate of the uncertain lifetime of bridges. For this, Jan and Leo Klatter proposed to use a Weibull distribution fitted to observed lifetimes of demolished bridges and censored lifetimes of existing bridges [27, 28] (later published in *Computers & Structures* [29]). The censored observations of the lifetimes were incorporated by using the left-truncated Weibull distribution.

With the information on the estimated lifetimes of bridges and the costs of various types of repairs, they defined a decision-theoretic approach to bridge management in the Netherlands [30, 31, 32, 33] (later published in *Structure and Infrastructure Engineering* [34]). Jan also looked into the application of the *Life-Quality Index* (LQI) for objectively assessing the increase in the quality of life in the Netherlands as a result of bridge maintenance [35]. Although this approach looked promising, it didn't really catch on in the bridge management community.

It is through his work for the Centre of Public Works that Jan met professor Dan Frangopol at the *International Conference on Structural Faults and Repairs* held in London in 1999, where they agreed to colaborate on research in the area of maintenance modeling. In particular, they compared the LEM model with the time-dependent reliability models developed by Dan Frangopol and his co-workers; see [36], which was later published in *Probabilistic Engineering Mechanics* [37]. In 2004, they published an invited paper, together with Maarten-Jan Kallen, with a review of probabilistic models for structural performance [38].

Maarten-Jan Kallen started his Ph.D. research under the supervision of Jan van Noortwijk in April 2003. Jan arranged for him to be an employee at HKV Consultants and to be hired as a consultant by the Centre for Public Works. It is a typical example of his ability to connect scientific research with business and it is reminiscent of the collaboration between the university in Delft and Delft Hydraulics during his own Ph.D. research. Before this time, Jan had already supervised Maarten-Jan during his M.Sc. project, which applied the gamma process for modeling the deterioration in pressure vessels used by the oil and gas industry. Companies which operate

these types of vessels are increasingly turning to a more probabilistic approach, known as 'Risk-Based Inspections' (RBI), for planning their inspections. The main results of his M.Sc. thesis were published in a paper at the ESREL conference in Maastricht, the Netherlands in 2003 [39], which later appeared in a special issue of the journal *Reliability Engineering and System Safety* [40]. Whereas these concerned the updating of the gamma process with information obtained using imperfect inspections, they also presented a paper, which considered multiple failure modes in the maintenance optimization of these pressure vessels, at the joint ESREL and PSAM conference in Berlin in 2004 [41]. This was not done by considering a bivariate deterioration process, but by reformulating the probabilities of preventive and corrective replacements due to at least one of these failure modes. This particular approach assumes that both degradation processes are independent.

The original idea for Maarten-Jan's Ph.D. project was to apply the gamma process for modeling bridge deterioration, but it soon became clear that insufficient data was available for this purpose. The Centre did have a database with data from visual inspections performed over a period of more than 20 years. It is therefore that the focus of the research shifted to fitting finite-state Markov processes to this data by use of appropriate methods for estimating the rate of transitions between condition states. The results of this research were published in papers at the ESREL conference held in Poland in 2005 [42], the IABMAS conference held in Portugal in 2006 [43], and in a special issue of the *International Journal of Pressure Vessels and Piping* [44] for which the model was reformulated to fit into the context of pressure vessels.

4.3 Sewer system management

Jan's first Ph.D. student was Hans Korving, who performed his research at HKV and at the section of Sanitary Engineering at the faculty of Civil Engineering and Geosciences of the Delft University of Technology. His supervisor there was prof. François Clement. Hans did his research towards the probabilistic modeling of the hydraulic performance and the management of the operational and structural condition of sewer systems. The overall aim was to include uncertainties of various types when making decision concerning the design, operation and maintenance of sewer systems [45, 46]. They used Bayesian statistics to determine the return period of combined sewer overflow (CSO) volumes, which is information that can be used for the risk-based design of such systems [47, 48]. For the maintenance and reliability modeling of sewer systems, they analysed failure data of sewage pumps assuming a non-homogeneous Poisson process for the occurrence of failures [49]. They also proposed a Bayesian model for updating prior knowledge on the condition state of sewer systems with the results of visual inspections [50]. The work presented in this paper is related to the work that Jan did for his M.Sc. thesis [2]. In the Netherlands, the condition of sewer systems is classified in one of five states according to the provisions

by the European norm NEN-EN-13508-2. If the likelihood of being in one these states is represented by a multinomial distribution, then the Dirichlet distribution may be used as a conjugate prior.

4.4 Corrosion modeling

In the publications [39, 40, 41] with Maarten-Jan Kallen, Jan van Noortwijk considered the thinning of steel walls due to corrosion and the process of stress-corrosion cracking. Using a gamma process to model the uncertain rate of thinning and cracking, they proposed a model which is updated with the results of imperfect (i.e., inaccurate) inspections. At the Centre for Public Works, Jan also considered problems related to corrosion of steel structures. Using the LEM model, he compared different strategies for the maintenance of the coating on the steel doors in the 'Haringvliet' storm-surge barrier [51]. He also co-authored a survey on deterioration models for corrosion modeling [52] together with Robin Nicolai and his Ph.D. supervisor at the time, Rommert Dekker.

Jan also published a few papers together with another Ph.D. student, Sebastian Kuniewski, whose research is sponsored by Shell Global Solutions in Amsterdam. His research is primarily focused on corrosion modeling of steel pipelines and vessels. In particular, they consider a form of sampling inspection, which is performed when a complete inspection of the whole surface of an object is not feasible. The information obtained from this partial inspection is then used to estimate the distribution of the largest defects in those areas which were not inspected [53, 54]. In a paper together with a former M.Sc. student of Jan, Juliana López de la Cruz, they looked at identifying clusters of pit corrosion in steel [55], based on a method to assess the goodness-of-fit of a non-homogeneous Poisson point process.

4.5 Gamma processes and renewal theory

Jan van Noortwijk is possibly best known for his work on the use of gamma processes for the stochastic modeling of deterioration. Starting with his Ph.D. thesis and ending with a survey of the application of gamma processes in maintenance [56] (published in *Reliability Engineering and System Safety* after his death in 2009), he published many papers in various subject areas in which the gamma process was used to model continuous and monotonically increasing processes of deterioration. Some variations included the combined probability of failure due to wear and randomly occuring schocks [57] (later published in a special issue of *Reliability Engineering and System Safety* [58]) and a bivariate gamma process to model two dependent deterioration processes [59].

Many of these publications were co-authored by prof. Mahesh Pandey from the University of Waterloo in Canada. Together with Hans van der Weide from the Delft University of Technology, he travelled to Canada for extended periods of time on several occasions, and they were in the process of writing a book together. Together with Mahesh, Jan published several

papers which were aimed at 'promoting' the use of the gamma process in the area of civil engineering. At three different conferences, they presented similar papers which highlighted the benefits of using the gamma process: an IFIP working conference in 2003 [60] (Jan has also written a general paper on the use of the gamma process for condition-based maintenance optimization at an earlier IFIP conference [61]), the IABMAS conference in 2004 [62], and the ICOSSAR conference in 2005 [63]. Finally, the contents of these papers were also published in *Structure and Infrastructure Engineering* in 2009 [64].

Another topic Jan worked on together with Hans and Mahesh, is the use of renewal theory in maintenance and reliability. In particular, they worked on various forms of monetary discounting for comparing future streams of expenditures based on their present value [65, 66, 67, 68]. This research followed Jan's work on cost-based criteria for maintenance decisions, in which he also considered the variance of costs [69] (later published in *Reliability Engineering and System Safety* [70]). In most cases, the policy with the lowest expected costs is chosen, but these papers show that the costs of these policies have the highest uncertainty (i.e., the largest variance) associated with them. In [71] (later publised in *Reliability Engineering and System Safety* [72] and used in [73]).

During his professional career, Jan van Noortwijk became a respected consultant and researcher in the area of maintenance optimization and reliability modeling. His authority in these subject areas is confirmed by his position as professor at the Delft University of Technology, by his position as lecturer at courses organized by the Foundation for Post Graduate Education in Delft, and the numerous invited papers and articles for journals and encyclopedia. For the *Wiley Encyclopedia of Statistics in Quality and Reliability*, he co-authored two articles: one on models for stochastic deterioration [74] and one on maintenance optimization [75].

5 FLOOD RISK MANAGEMENT

Jan started his research in flood risk management in 1994 with an uncertainty analysis of strategies to reduce the risk of flooding in the river Meuse. This research was carried out in a Delft Hydraulics project for the Committee Flood Disaster Meuse [76]. It became the topic of the eighth chapter in his Ph.D. thesis and it later also became a chapter in the book *The practice of Bayesian Analysis* [77]. The new idea of his approach was to use a Bayesian approach for the assessment of the uncertainties in the expected flood damage and the costs of the strategies. The most important uncertainties were the river discharge, the flood damage given the discharge, the downstream water levels along the Meuse given the discharge, and the costs and benefits of decisions.

In one of his first projects at HKV, Jan derived the generalised gamma distribution for modelling the uncertain size of peak discharges in the Rhine

river [78]. This particular probability distribution has the advantage of fitting well with the stage-discharge curve being an approximate power law between water level and discharge [79].

In 1996, Jan made a big contribution in a study on the modeling of the roughness of submerged vegetation [80]. A new, analytical, physics-based model of the vertical flow velocity profile and the hydraulic roughness of submerged vegetation was developed. Jan found an analytical solution to the differential equation of the model, which was not known in the literature at that time and which opened a wide range of applications. Another contribution in this area is the calibration of hydraulic models. Calibration of these mathematical models is a time consuming process. This process can be automated by function minimisation with the simplex algorithm. In [81] it is described how Jan, together with two colleagues (Matthijs Duits and Anne Wijbenga), contributed to this problem with an application to one of the Dutch rivers.

The contributions of Jan in the field of flood risk management were remarkable and included an amazing number of topics. In particular, he covered both aspects of risk, namely the probability of occurrence and the consequences of a flood event. In the following sections, an overview of his contributions to both aspects is given.

5.1 The probability of occurrence of a flood

The main contribution of Jan van Noortwijk in flood risk management has been the use of Bayesian statistics. Jan has written nine papers about this topic [79, 82, 83, 84, 85, 86, 87, 88, 89] and he also initiated a common research program between HKV Consultants and the Ministry of Transport, Public Works and Water Management, from 2000 to 2008. Program leader on behalf of the Ministry was mr. Houcine Chbab. This research program resulted in new Bayesian methods and a software program to apply these methods in practice. One of the applications is the assessment of 'design' discharges of rivers, which represent the discharges with a given return period (i.e., the reciprocal of the probability of exceedance). In the classical approach, statistical uncertainties are not taken into account. In the Bayesian approach, the prior distribution represents information about the uncertainty of the statistical parameters, and, using Bayes' theorem, it can be updated with the available data. So, rather than choosing one particular probability distribution a priori, Jan proposed to fit various probability distributions to the observations and to attach weights to these distributions according to how well they fit this data. So-called Bayes factors are used to determine these weights. Another major contribution is his derivation of non-informative Jeffrey's priors for a large number of probability distributions. Data from many rivers (for example, the Rhine and Oder rivers) and results of the Bayesian approach are included in the papers. An important conclusion is that the design discharges increase when taking the statistical uncertainties into account properly [88].

Information on water levels and discharges is important in order to determine the probability of the failure mode of 'overtopping' in which the waterlevel exceeds the crest-level of a dike. In [90], a special Monte Carlo method (directional sampling) was used to assess the probability of dike failure due to the failure mechanism 'uplifting and piping'. Dike failure due to uplifting and piping is defined as the event in which the resistance (the critical head) drops below the stress (the outer water level minus the inner water level). Special attention was given to the spatial variation, since the critical head is correlated over the length of a dike. The correlation is modelled using a Markovian dependency structure. The paper shows results of a dike section in the lower river area of the Netherlands.

5.2 The consequences of a flood

Jan also made extensive use of the methods developed by Roger Cooke in the field of expert judgment. In his Master's thesis, Jan elicited expert opinions on reliability data in a structured manner. In 2005, he formulated a new method for determining the time available for evacuation of a dike-ring area by expert judgment [91]. This research was done together with HKV colleague Anne Barendregt and two experts from the Ministry of PublicWorks and Water Management: Stephanie Holterman and Marcel van der Doef. They addressed the following problem. The possibilities open to preventive evacuation because of a flood threat depend on the time available and the time required for evacuation. If the time available for evacuation is less than the time required, complete preventive evacuation of an area is not possible. Because there are almost no observations on the time available, Jan and his colleagues had to rely on expert opinions. It is remarkable that the results of this study are still of value. It is widely recognized that the methodology was sound, and that the expert elicitation was done with much care.

Together with Anne Barendregt, Stephanie Holterman and an M.Sc. student from Delft, Regina Egorova, Jan published results on an effort to quantify the uncertainty in flood damage estimation [92]. They considered uncertainty in the maximum damage per object and the damage function. Given the water level, the damage function gives the damage incurred as a fraction of the maximum damage. The uncertainty in the damage function was represented by a Beta distribution. Finally, they also considered the effect of spatial dependence between the damages in a flooded area and they applied the model to the Central-Holland dike-ring area.

5.3 Cost-benefit analysis of flood protection measures

The area of cost-benefit analysis of measures for flood protection was also covered by Jan. In [8], he addressed the problem of how to achieve cost-optimal dike heightening for which the sum of the initial cost of investment and the future (discounted) cost of maintenance is minimal. Jan developed a maintenance model for dikes subject to uncertain crest-level decline. On the basis of engineering knowledge, crest-level decline was modeled as

a monotone stochastic process with expected decline being linear of non-linear in time. For a particular unit of time, the increments are distributed according to mixtures of exponentials. In a case study, the maintenance decision model has been applied to the problem of heightening the Dutch 'Oostmolendijk'. In [57, 58], Jan addressed the time dependent reliability of the Den Helder sea defence as stochastic processes of deteriorating resistance and hydraulic load. Recently, Jan also addressed the cost-benefit method of flood protection as a non-stationary control problem, as suggested by mr. Carel Eigenraam of the Central Planning Office. Here, the benefits of a decision are modeled as the present value of expected flood damage. Jan has written two HKV reports about this optimization problem, and also guided one of his M.Sc. students, Bastiaan Kuijper, in this direction (this research was recenty published as [93]). Unfortunately, he was unable to enrich the scientific literature with more publications on this topic.

Acknowledgments

Many people have given valuable input during the writing of this overview and for which we are very grateful. In particular, we would like to thank Roger Cooke, Rommert Dekker, René van Dorp, Pieter van Gelder, Jaap Bakker, Hans Korving, Thomas Mazzuchi and Hans van der Weide for their help in putting many of the publications by Jan van Noortwijk in a historical context. Any omissions or inaccuracies of facts presented in this overview are solely due to the authors.

Bibliography

[1] Rommert Dekker and Cyp F. H. van Rijn. PROMPT - a decision support system for opportunity based preventive maintenance. In S. Ozekici, editor, *Reliability and Maintenance of Complex Systems*, volume 154 of *NATO ASI series*, pages 530–549. Springer-Verlag, Berlin, 1996.

[2] J. M. van Noortwijk. Use of expert opinion for maintenance optimisation. Master's thesis, Delft University of Technology, The Netherlands, 1989.

[3] J. M. van Noortwijk, R. Dekker, R. M. Cooke, and T. A. Mazzuchi. Expert judgment in maintenance optimization. *IEEE Transactions on Reliability*, 41(3):427–432, 1992.

[4] T. A. Mazzuchi, R. M. Cooke, J. R. van Dorp, J. M. van Noortwijk, and R. Dekker. The elicitation and use of expert judgment for maintenance optimization. In Jay Liebowitz, editor, *The World Congress on Expert Systems, Orlando, Florida, Volume 2*, pages 888–896, 1991.

[5] V. Dijk, E. Aanen, H. van den Berg, and J. M. van Noortwijk. Extrapolating atm-simulation results using extreme value theory. In J. W. Cohen and C. D. Pack, editors, *Queueing, Performance and Control in ATM (13th International Teletraffic Congress, Copenhagen)*, pages 97–104, Amsterdam, 1991. Elsevier Science Publishers B. V.

[6] J. M. van Noortwijk. *Optimal Maintenance Decisions for Hydraulic Structures under Isotropic Deterioration*. PhD thesis, Delft University of Tech-

nology, The Netherlands, 1996.

[7] J. M. van Noortwijk and E. B. Peerbolte. Optimal sand nourishment decisions. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 126(1): 30–38, 2000.

[8] L. J. P. Speijker, J. M. van Noortwijk, M. Kok, and R. M. Cooke. Optimal maintenance decisions for dikes. *Probability in the Engineering and Informational Sciences*, 14(4):101–121, 2000.

[9] J. M. van Noortwijk and P. H. A. J. M. van Gelder. Optimal maintenance decisions for berm breakwaters. *Structural Safety*, 18(4):293–309, 1996.

[10] J. M. van Noortwijk and H. E. Klatter. Optimal inspection decisions for the block mats of the Eastern-Scheldt barrier. *Reliability Engineering and System Safety*, 65(3):203–211, 1999.

[11] J. M. van Noortwijk, M. Kok, and R. M. Cooke. Optimal maintenance decisions for the sea-bed protection of the Eastern-Scheldt barrier. In R. Cooke, M. Mendel, and H. Vrijling, editors, *Engineering Probabilistic Design and Maintenance for Flood Protection*, pages 25–56. Dordrecht: Kluwer Academic Publishers, 1997.

[12] Max Bernhard Mendel. *Development of Bayesian Parametric Theory with Applications to Control*. PhD thesis, Massachusetts Institute of Technology, U.S.A., 1989.

[13] R. E. Barlow and M. B. Mendel. De Finetti-type representations for life distributions. *Journal of the American Statistical Association*, 87(420):1116–1122, 1992.

[14] J. M. van Noortwijk. Inspection and repair decisions for hydraulic structures under symmetric deterioration. Technical Report ESRC 92-17, University of California at Berkeley, U.S.A., 1992.

[15] J. M. van Noortwijk, R. M. Cooke, and M. Kok. A Bayesian failure model based on isotropic deterioration. *European Journal of Operational Research*, 82(2):270–282, 1995.

[16] J. M. van Noortwijk, R. M. Cooke, and J. K. Misiewicz. Characterisations of scale mixtures of gamma processes in terms of sufficiency and isotropy. *Journal of Mathematical Sciences*, 99(4):1469–1475, 2000.

[17] J. M. van Noortwijk. Optimal maintenance decisions on the basis of uncertain failure probabilities. *Journal of Quality in Maintenance Engineering*, 6 (2):113–122, 2000.

[18] J. M. van Noortwijk. Preventief onderhoud van waterbouwkundige constructies; probleeminventarisatie [Preventive maintenance of hydraulic structures; inventory of problems]. Technical Report Q1210, Delft Hydraulics, The Netherlands, 1991.

[19] J. J. de Jonge, M. Kok, and J. M. van Noortwijk. Onderhoud van de natte werken van waterbouwkundige constructies [Maintenance of the wet works of hydraulic structures]. Technical Report Q1297, Delft Hydraulics, The Netherlands, 1991.

[20] J. M. van Noortwijk and J. H. de Jong. Lay-outs and ship-passage conditions for the "Nieuwe Stadsbrug" pairwise compared; an application of the method of paired comparisons. Technical Report Q1311, Delft Hydraulics, The Netherlands, 1992.

[21] R. M. Cooke and J. M. van Noortwijk. Local probabilistic sensitivity measures for comparing FORM and Monte Carlo calculations illustrated with

dike ring reliability calculations. *Computer Physics Communications*, 117 (1–2):86–98, 1999.

[22] R. M. Cooke and J. M. van Noortwijk. Generic graphics for uncertainty and sensitivity analysis. In Schuëller and Kafka [102], pages 1187–1192.

[23] R. M. Cooke and J. M. van Noortwijk. Graphical methods. In A. Saltelli, K. Chan, and E. M. Scott, editors, *Sensitivity Analysis*, pages 245–264, Chichester, 2000. John Wiley & Sons.

[24] J. D. Bakker, H. J. van der Graaf, and J. M. van Noortwijk. Model of lifetime-extending maintenance. In M. C. Forde, editor, *Proceedings of the 8th International Conference on Structural Faults and Repair*, Edinburgh, July 1999. Engineering Technics Press.

[25] A. van Beek, G. C. M. Gaal, J. M. van Noortwijk, and J. D. Bakker. Validation model for service life prediction of concrete structures. In D. J. Naus, editor, *2nd International RILEM Workshop on Life Prediction and Aging Management of Concrete Structures, 5-6 May 2003, Paris, France*, pages 257–267, Bagneux, 2003. International Union of Laboratories and Experts in Construction Materials, Systems and Structures (RILEM).

[26] J. D. Bakker and J. M. van Noortwijk. Inspection validation model for life-cycle analysis. In Watanabe et al. [94].

[27] J. M. van Noortwijk and H. E. Klatter. The use of lifetime distributions in bridge replacement modelling. In Casas et al. [95].

[28] H. E. Klatter and J. M. van Noortwijk. Life-cycle cost approach to bridge management in the Netherlands. In *Proceedings of the 9th International Bridge Management Conference, April 28-30, 2003, Orlando, Florida, U.S.A., Transportation Research Circular E-C049*, pages 179–188, Washington D.C., 2003. Transportation Research Board (TRB).

[29] J. M. van Noortwijk and H. E. Klatter. The use of lifetime distributions in bridge maintenance and replacement modelling. *Computers and Structures*, 82(13–14):1091–1099, 2004.

[30] H. E. Klatter, J. M. van Noortwijk, and N. Vrisou van Eck. Bridge management in the Netherlands; prioritisation based on network performance. In Casas et al. [95].

[31] H. E. Klatter, A. C. W. M. Vrouwenvelder, and J. M. van Noortwijk. Life-cycle-cost-based bridge management in the Netherlands. In Watanabe et al. [94].

[32] H. E. Klatter, A. C. W. M. Vrouwenvelder, and J. M. van Noortwijk. Societal aspects of bridge management and safety in the Netherlands. In Cruz et al. [96].

[33] H. E. Klatter and J. M. van Noortwijk. Lcc analysis of structures on a network level in the Netherlands. In H-N. Cho, D. M. Frangopol, A. H-S. Ang, and J. S. Kong, editors, *Life-Cycle Cost and Performance of Civil Infrastructure Systems, Proceedings of the Fifth International Workshop on Life-Cycle Cost Analysis and Design of Civil Infrastructure Systems, Seoul, Korea, 16-18 October 2006*, pages 215–220, London, 2007. Taylor & Francis Group.

[34] H. E. Klatter, T. Vrouwenvelder, and J. M. van Noortwijk. Societal and reliability aspects of bridge management in the netherlands. *Structure and Infrastructure Engineering*, 5(1):11–24, 2009.

[35] M. D. Pandey, J. M. van Noortwijk, and H. E. Klatter. The potential

applicability of the life-quality index to maintenance optimisation problems. In Cruz et al. [96].

[36] J. M. van Noortwijk and D. M. Frangopol. Deterioration and maintenance models for insuring safety of civil infrastructures at lowest life-cycle cost. In D. M. Frangopol, E. Brühwiler, M. H. Faber, and B. Adey, editors, *Life-Cycle Performance of Deteriorating Structures: Assessment, Design and Management*, pages 384–391, Reston, Virginia, 2004. American Society of Civil Engineers (ASCE).

[37] J. M. van Noortwijk and D. M. Frangopol. Two probabilistic life-cycle maintenance models for deteriorating civil infrastructures. *Probabilistic Engineering Mechanics*, 19(4):345–359, 2004.

[38] D. M. Frangopol, M. J. Kallen, and J. M. van Noortwijk. Probabilistic models for life-cycle performance of deteriorating structures: review and future directions. *Progress in Structural Engineering and Materials*, 6(4): 197–212, 2004.

[39] M. J. Kallen and J. M. van Noortwijk. Inspection and maintenance decisions based on imperfect inspections. In Bedford and van Gelder [97], pages 873–880.

[40] M. J. Kallen and J. M. van Noortwijk. Optimal maintenance decisions under imperfect inspection. *Reliability Engineering and System Safety*, 90 (2–3):177–185, 2005.

[41] M. J. Kallen and J. M. van Noortwijk. Optimal inspection and replacement decisions for multiple failure modes. In C. Spitzer, U. Schmocker, and V. N. Dang, editors, *Probabilistic Safety Assessment and Management (PSAM7-ESREL'04): Proceedings of the 7th International Conference on Probabilistic Safety Assessment and Management, 14-18 June Berlin, Germany*, pages 2435–2440, London, 2004. Springer-Verlag.

[42] M. J. Kallen and J. M. van Noortwijk. A study towards the application of Markovian deterioration processes for bridge maintenance modelling in the Netherlands. In Ko lowrocki [98], pages 1021–1028.

[43] M. J. Kallen and J. M. van Noortwijk. Statistical inference for markov deterioration models of bridge conditions in the Netherlands. In Cruz et al. [96].

[44] M. J. Kallen and J. M. van Noortwijk. Optimal periodic inspection of a deterioration process with sequential condition states. *International Journal of Pressure Vessels and Piping*, 83(4):249–255, 2006.

[45] H. Korving, J. M. van Noortwijk, P. H. A. J. M. van Gelder, and F. L. H. R. Clemens. Influence of model parameter uncertainties on decision-making for sewer system management. In I. D. Cluckie, D. Han, J. P. Davis, and S. Heslop, editors, *Proceedings of the Fifth International Conference on Hydroinformatics, July 1-5, 2002, Cardiff, United Kingdom; Volume 2: Software Tools and Management Systems*, pages 1361–1366, London, 2002. International Water Association (IWA) Publishing.

[46] H. Korving, J. M. van Noortwijk, P. H. A. J. M. van Gelder, and R. S. Parkhi. Coping with uncertainty in sewer system rehabilitation. In Bedford and van Gelder [97], pages 959–967.

[47] H. Korving, F. Clemens, J. van Noortwijk, and P. van Gelder. Bayesian estimation of return periods of CSO volumes for decision-making in sewer system management. In E. W. Strecker and W. C. Huber, editors, *Global So-*

lutions for Urban Drainage, Proceedings of the Nineth International Conference on Urban Drainage, September 8-13, 2002, Portland, Oregon, U.S.A.., New York, 2002. American Society of Civil Engineers (ASCE).

[48] H. Korving, J. M. van Noortwijk, P. H. A. J. M. van Gelder, and F. H. L. R. Clemens. Risk-based design of sewer system rehabilitation. *Structure and Infrastructure Engineering*, 5(3):215–227, 2009.

[49] H. Korving, F. H. L. R. Clemens, and J. M. van Noortwijk. Statistical modeling of the serviceability of sewage pumps. *Journal of Hydraulic Engineering*, 132(10):1076–1085, 2006.

[50] H. Korving and J. M. van Noortwijk. Bayesian updating of a prediction model for sewer degradation. In T. Ertl, A. Pressl, F. Kretschmer, and R. Haberl, editors, *Proceedings of the Second International IWA Conference on Sewer Operation and Maintenance, 26-28 October 2006, Vienna, Austria*, pages 199–206, Vienna, Austria, 2006. Institute of Sanitary Engineering and Water Polution Control (BOKU).

[51] A. Heutink, A. van Beek, J. M. van Noortwijk, H. E. Klatter, and A. Barendregt. Environment-friendly maintenance of protective paint systems at lowest costs. In *XXVII FATIPEC Congres; 19-21 April 2004, Aix-en-Provence*, pages 351–364. AFTPVA, Paris, 2004.

[52] R. P. Nicolai, R. Dekker, and J. M. van Noortwijk. A comparison of models for measurable deterioration: an application to coatings on steel structures. *Reliability Engineering and System Safety*, 92(12):1635–1650, 2007.

[53] S. P. Kuniewski and J. M. van Noortwijk. Sampling inspection for the evaluation of time-dependent reliability of deteriorating structures. In Aven and Vinnem [99], pages 281–288.

[54] S. P. Kuniewski, J. A. M. van der Weide, and J. M. van Noortwijk. Sampling inspection for the evaluation of time-dependent reliability of deteriorating systems under imperfect defect detection. *Reliability Engineering & System Safety*, 94(9):1480–1490, 2009.

[55] J. López De La Cruz, S. P. Kuniewski, J. M. van Noortwijk, and M. A. Gutiérrez. Spatial nonhomogeneous poisson process in corrosion management. *Journal of The Electrochemical Society*, 155(8):C396–C406, 2008.

[56] J. M. van Noortwijk. A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1):2–21, 2009.

[57] J. M. van Noortwijk, M. J. Kallen, and M. D. Pandey. Gamma processes for time-dependent reliability of structures. In Ko lowrocki [98], pages 1457–1464.

[58] J. M. van Noortwijk, J. A. M. van der Weide, M. J. Kallen, and M. D. Pandey. Gamma processes and peaks-over-threshold distributions for time-dependent reliability. *Reliability Engineering and System Safety*, 92(12): 1651–1658, 2007.

[59] F. A. Buijs, J. W. Hall, J. M. van Noortwijk, and P. B. Sayers. Time-dependent reliability analysis of flood defences using gamma processes. In Augusti et al. [100], pages 2209–2216.

[60] J. M. van Noortwijk and M. D. Pandey. A stochastic deterioration process for time-dependent reliability analysis. In M. A. Maes and L. Huyse, editors, *Proceedings of the Eleventh IFIP WG 7.5 Working Conference on Reliability and Optimization of Structural Systems, Banff, Canada, 2-5 November 2003*, pages 259–265, London, 2004. Taylor & Francis Group.

[61] J. M. van Noortwijk. Optimal replacement decisions for structures under stochastic deterioration. In A. S. Nowak, editor, *Proceedings of the Eighth IFIP WG 7.5 Working Conference on Reliability and Optimization of Structural Systems, Kraków, Poland, 11-13 May 1998*, pages 273–280, Ann Arbor, 1998. University of Michigan.

[62] M. D. Pandey and J. M. van Noortwijk. Gamma process model for time-dependent structural reliability analysis. In Watanabe et al. [94].

[63] M. D. Pandey, X.-X. Yuan, and J. M. van Noortwijk. Gamma process model for reliability analysis and replacement of aging structural components. In Augusti et al. [100], pages 2209–2216.

[64] M. D. Pandey, X.-X. Yuan, and J. M. van Noortwijk. The influence of temporal uncertainty of deterioration in life-cycle management of structures. *Structure and Infrastructure Engineering*, 5(2):145–156, 2009.

[65] J. A. M. van der Weide, J. M. van Noortwijk, and Suyono. Application of probability in constructing dykes. *Jurnal Matematika, Statistika dan Komputasi*, pages 1–9, 2007.

[66] J. A. M. van der Weide, J. M. van Noortwijk, and Suyono. Renewal theory with discounting. In *Proceedings of the Fifth MMR (Mathematical Methods in Reliability) Conference 2007, Glasgow, Scotland, July 1-4, 2007, CD-ROM*, Glasgow, 2007. University of Strathclyde.

[67] J. A. M. van der Weide, Suyono, and J. M. van Noortwijk. Renewal theory with exponential and hyperbolic discounting. *Probability in the Engineering and Informational Sciences*, 22(1):53–74, 2008.

[68] J. A. M. van der Weide, J. M. van Noortwijk, and Suyono. Renewal theory with discounting. In T. Bedford, J. Quigley, L. Walls, B. Alkali, A. Daneshkah, and G. Hardman, editors, *Advances in Mathematical Modeling for Reliability*, Netherlands, 2008. IOS Press.

[69] J. M. van Noortwijk. Cost-based criteria for obtaining optimal design decisions. In Corotis et al. [101].

[70] J. M. van Noortwijk. Explicit formulas for the variance of discounted life-cycle cost. *Reliability Engineering and System Safety*, 80(2):185–195, 2003.

[71] J. M. van Noortwijk and J. A. M. van der Weide. Computational techniques for discrete-time renewal processes. In C. Guedes Soares and E. Zio, editors, *Safety and Reliability for Managing Risk, Proceedings of ESREL 2006 – European Safety and Reliability Conference 2006, Estoril, Portugal, 18-22 September 2006*, pages 571–578, London, 2006. Taylor & Francis Group.

[72] J. M. van Noortwijk and J. A. M. van der Weide. Applications to continuous-time processes of computational techniques for discrete time renewal processes. *Reliability Engineering & System Safety*, 93(12):1853–1860, 2008.

[73] J. A. M. van der Weide, M. D. Pandey, and J. M. van Noortwijk. A conceptual interpretation of the renewal theorem with applications. In Aven and Vinnem [99], pages 477–484.

[74] J. M. van Noortwijk and M. J. Kallen. Stochastic deterioration. In F. Ruggeri, R. S. Kenett, and F. W. Faltin, editors, *Encyclopedia of Statistics in Quality and Reliability*, pages 1925–1931. John Wiley & Sons, Chichester, 2007.

[75] T. A. Mazzuchi, J. M. van Noortwijk, and M. J. Kallen. Maintenance optimization. In F. Ruggeri, R. S. Kenett, and F. W. Faltin, editors, *Encyclopedia of Statistics in Quality and Reliability*, pages 1000–1008. John

Wiley & Sons, Chichester, 2007.

[76] J. M. van Noortwijk and M. Kok. Onderzoek Watersnood Maas. Deelrapport 14: Onzekerheidsanalyse. [Investigation of the Meuse Flood. Subreport 14: Uncertainty Analysis]. Technical Report Q1858/T1349, Delft Hydraulics & Delft University of Technology, The Netherlands, 1994.

[77] J. M. van Noortwijk, M. Kok, and R. M. Cooke. Optimal decisions that reduce flood damage along the Meuse: an uncertainty analysis. In S. French and J. Q. Smith, editors, *The Practice of Bayesian Analysis*, pages 151–172, London, 1997. Arnold.

[78] M. Kok, N. Douben, J. M. van Noortwijk, and W. Silva. Integrale Verkenning inrichting Rijntakken – Veiligheid [River Management of the Rhine Branches – Safety]. Technical Report IVR nr. 12, Ministerie van Verkeer en Waterstaat [Ministry of Transport, Public Works and Water Management], the Netherlands, 1996.

[79] J. M. van Noortwijk. Bayes estimates of flood quantiles using the generalised gamma distribution. In Y. Hayakawa, T. Irony, and M. Xie, editors, *System and Bayesian Reliability: Essays in Honor of Professor Richard E. Barlow*, pages 351–374, Singapore, 2001. World Scientific Publishing.

[80] D. Klopstra, H. J. Barneveld, J. M. van Noortwijk, and E. H. van Velzen. Analytical model for hydraulic roughness of submerged vegetation. In F. M. Holly Jr. and A. Alsaffar, editors, *Water for A Changing Global Community, The 27th Congress of the International Association for Hydraulic Research, San Francisco, 1997; Proceedings of Theme A, Managing Water: Coping with Scarcity and Abundance*, pages 775–780, New York, 1997. American Society of Civil Engineers (ASCE).

[81] J. H. A. Wijbenga, M. T. Duits, and J. M. van Noortwijk. Parameter optimisation for two-dimensional flow modelling. In V. Babovic and L. C. Larsen, editors, *Proceedings of the Third International Conference on Hydroinformatics, Copenhagen, Denmark, 1998*, pages 1037–1042, Rotterdam, 1998. Balkema.

[82] P. H. A. J. M. van Gelder, J. M. van Noortwijk, and M. T. Duits. Selection of probability distributions with a case study on extreme Oder river discharges. In Schuëller and Kafka [102], pages 1475–1480.

[83] J. M. van Noortwijk and P. H. A. J. M. van Gelder. Bayesian estimation of quantiles for the purpose of flood prevention. In B. L. Edge, editor, *Proceedings of the 26th International Conference on Coastal Engineering, Copenhagen, Denmark, 1998*, pages 3529–2541, New York, 1999. American Society of Civil Engineers (ASCE).

[84] E. H. Chbab, J. M. van Noortwijk, and M. T. Duits. Bayesian frequency analysis of extreme river discharges. In F. Toensmann and M. Koch, editors, *River Flood Defence: Proceedings of the International Symposium on Flood Defence, Kassel, Germany, 2000*, pages F51–F60, Kassel, 2000. Herkules Verlag Kassel.

[85] E. H. Chbab, J. M. van Noortwijk, and H. J. Kalk. Bayesian estimation of extreme discharges. In M. Spreafico and R. Weingartner, editors, *CHR Report II-17, International Conference on Flood Estimation, March 6-8, 2002, Berne, Switzerland*, pages 285–294, Lelystad, 2002. International Commission for the Hydrology of the Rhine basin (CHR).

[86] J. M. van Noortwijk, H. J. Kalk, and E. H. Chbab. Bayesian computation

of design discharges. In T. Bedford and P. H. A. J. M. van Gelder, editors, *Proceedings of ESREL 2003 – European Safety and Reliability Conference '03, 15-18 June 2003, Maastricht, The Netherlands*, pages 1179–1187. Rotterdam: Balkema, 2003.

[87] J. M. van Noortwijk, H. J. Kalk, M. T. Duits, and E. H. Chbab. Bayesian statistics for flood prevention. Technical Report PR280, Ministry of Transport, Public Works and Water Management, Institute for Inland Water Management and Waste Water Treatment (RIZA), and HKV Consultants, Lelystad, The Netherlands, 2003.

[88] J. M. van Noortwijk, H. J. Kalk, M. T. Duits, and E. H. Chbab. The use of Bayes factors for model selection in structural reliability. In Corotis et al. [101].

[89] J. M. van Noortwijk, H. J. Kalk, and E. H. Chbab. Bayesian estimation of design loads. *HERON*, 49(2):189–205, 2004.

[90] J. M. van Noortwijk, A. C. W. M. Vrouwenvelder, E. O. F. Calle, and K. A. H. Slijkhuis. Probability of dike failure due to uplifting and piping. In Schuëller and Kafka [102], pages 1165–1170.

[91] A. Barendregt, J. M. van Noortwijk, M. van der Doef, and S. R. Holterman. Determining the time available for evacuation of a dike-ring area by expert judgement. In J. K. Vrijling, E. Ruijgh, B. Stalenberg, P. H. A. J. M. van Gelder, M. Verlaan, A. Zijderveld, and P. Waarts, editors, *Proceedings of the Nineth International Symposium on Stochastic Hydraulics (ISSH), Nijmegen, The Netherlands, 23-23 May 2005, pages on CD-ROM*, Madrid, 2005. International Association of Hydraulic Engineering and Research (IAHR).

[92] R. Egorova, J. M. van Noortwijk, and S. R. Holterman. Uncertainty in flood damage estimation. *International Journal of River Basin Management*, 6 (2):139–148, 2008.

[93] B. Kuijper and M. J. Kallen. The impact of risk aversion on optimal economic decisions. In R. Bris, C. Guedes Soares, and S. Martorell, editors, *Reliability, Risk and Safety: Theory and Applications, Proceedings of the European Safety and Reliability Conference (ESREL), Prague, September 7-10, 2009*, volume 1, pages 453–460, London, 2010. Taylor & Francis Group.

[94] E. Watanabe, D. M. Frangopol, and T. Utsonomiya, editors. *Bridge Maintenance, Safety, Management and Cost, Proceedings of the Second International Conference on Bridge Maintenance, Safety and Management (IABMAS)*, Kyoto, Japan, 18-22 October 2004, 2004. Taylor & Francis Group, London.

[95] J. R. Casas, D. M. Frangopol, and A. S. Nowak, editors. *First International Conference on Bridge Maintenance, Safety and Management (IABMAS)*, Barcelona, Spain, 14-17 July 2002, 2002. International Center for Numerical Methods in Engineering (CIMNE).

[96] P. J. S. Cruz, D. M. Frangopol, and L. C. Neves, editors. *Bridge Maintenance, Safety, Management, Life-Cycle Performance and Cost, Proceedings of the Third International Conference on Bridge Maintenance, Safety and Management (IABMAS)*, Porto, Portugal, 16-19 July 2006, CD-ROM, 2006. Taylor & Francis Group, London.

[97] T. Bedford and P. H. A. J. M. van Gelder, editors. *Proceedings of ESREL 2003 – European Safety and Reliability Conference '03*, 15-18 June 2003,

Maastricht, The Netherlands, 2003. Balkema, Rotterdam.

[98] K. Ko lowrocki, editor. *Advances in Safety and Reliability, Proceedings of ESREL 2005 – European Safety and Reliability Conference 2005*, Tri City (Gdynia-Sopot-Gdańsk), Poland, 27-30 June 2005, 2005. Taylor & Francis Group, London.

[99] T. Aven and J. E. Vinnem, editors. *Risk, Reliability and Societal Safety, Proceedings of ESREL 2007 – European Safety and Reliability Conference 2007,*, Stavanger, Norway, 25-27 June 2007, 2007. Taylor & Francis Group, London.

[100] G. Augusti, G. I. Schuëller, and M. Ciampoli, editors. *Safety and Reliability of Engineering Systems and Structures; Proceedings of the Nineth International Conference on Structural Safety and Reliability (ICOSSAR)*, Rome, Italy, 19-23 June 2005, 2005. Millpress, Rotterdam.

[101] R. B. Corotis, G. I. Schuëller, and M. Shinozuka, editors. *Structural Safety and Reliability – Proceedings of the Eighth International Conference on Structural Safety and Reliability (ICOSSAR)*, Newport Beach, California, U.S.A., 17-22 June 2001, 2001. Balkema, Lisse.

[102] G. I. Schuëller and P. Kafka, editors. *Safety and Reliability, Proceedings of ESREL Ś99 – The Tenth European Conference on Safety and Reliability, Munich-Garching, Germany, 1999*, Munich-Garching, Germany, 1999, 1999. Balkema, Rotterdam.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

On some elicitation procedures for distributions with bounded support – with applications in PERT

Johan René van Dorp*
– The George Washington University, Washington D.C., USA

Abstract. The introduction of the Project Evaluation and Review Technique (PERT) dates back to the 1960's and has found wide application since then in the planning of construction projects. Difficulties with the interpretation of the parameters of the beta distribution let Malcolm et al. [1] to suggest the classical expressions for the PERT mean and variance for activity completion that follow from lower and upper bound estimates a and b and a most likely estimate θ thereof. The parameters of the beta distribution are next estimated via the method of moments technique. Despite more recent papers still questioning the PERT mean and variance approach, their use is still prevalent in operations research and industrial engineering text books that discuss these methods. In this paper an overview is presented of some alternative approaches that have been suggested, including a recent approach that allows for a direct model range estimation combined with an indirect elicitation of bound and tail parameters of generalized trapezoidal uniform distributions describing activity uncertainty. Utilizing an illustrative Monte Carlo analysis for the completion time of an 18 node activity network, we shall demonstrate a difference between project completion times that could result when requiring experts to specify a single most likely estimate rather than allowing for a modal range specification.

1 INTRODUCTION

The three parameter triangular distribution $Triang(a, \theta, b)$, with lower and upper bounds a and b and most likely value θ, is one of the first continuous distributions on the bounded range proposed back in 1755 by English mathematician Thomas Simpson [2, 3]. It received special attention as late as in the 1960's, in the context of the PERT (see, e.g., Winston [4]) as an alternative to the four-parameter beta distribution:

$$f_T(t|a, b; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{(t-a)^{\alpha-1}(b-t)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}, \tag{1}$$

*corresponding author: Department of Engineering Management and Systems Engineering, School of Engineering and Applied Science, The George Washington University, 1776 G Street, N.W., Washington D.C. 20052, U.S.A.; telephone: +1-(202) 994 6638, fax: +1-(202) 994 0245, e-mail: dorpjr@gwu.edu.

with $a \leq t \leq b$, $\alpha > 0$, and $\beta > 0$. This distribution involves some difficulties regarding the interpretation of its parameters α and β. As a result, Malcolm et al. [1] suggested the following PERT mean and variance expressions

$$E[T] = \frac{a + 4\theta + b}{6},\ Var[T] = \frac{1}{36}(b-a)^2 \tag{2}$$

where T is a random variable modeling activity completion time, a and b being the lower and upper bound estimates and θ being a most like estimate for T. The remaining beta parameters α and β in (1) are next obtained from (2) utilizing the method of moments. Kamburowski [5] notes that: "Despite the criticisms and the abundance of new estimates, the PERT mean and variance (given by Equation (2) in this paper) can be found in almost every textbook on OR/MS and P/OM, and are employed in much project management software."

The somewhat non-rigorous proposition (2) resulted in a vigorous debate over 40 years ago (Clark [6], Grubbs [7], Moder and Rodgers [8]) regarding its appropriateness and even serves as the topic of more recent papers (see, e.g., Herrerías [9], Kamburowski [5], Herrerías et al. [10]). In a further response to the criticism of (2), Herrerías [9] suggested substitution of

$$\alpha = 1 + s(\theta - a)/(b-a), \beta = 1 + s(b-\theta)/(b-a), \tag{3}$$

in (1) instead, where $s > -1$ and $a < \theta < b$. This yields

$$E[T] = \frac{a + s\theta + b}{s+2},\ Var[T] = \frac{(s+1)(b-a)^2 + s^2(b-\theta)(\theta - a)}{(s+3)(s+2)^2}. \tag{4}$$

Essentially, Herrerías [9] reparameterizes the beta probability density function (PDF) in Equation (1) by managing to express α and β in terms of new parameters θ and s while retaining the lower and upper bounds a and b. For $s > 0$ the beta PDF (1) is unimodal and for $s = 0$ it reduces to a uniform distribution. Hence, Herrerías [9] designated s to be a confidence parameter in the mode location θ such that higher values of s indicate a higher confidence. Indeed, for $s \to \infty$, the beta pdf converges to a single point mass at θ. For $-1 < s < 0$, the beta PDF (11) is U-shaped which is not consistent with θ being a most likely value.

As a further alternative to the beta PDF (1), Van Dorp and Kotz [11] generalized the Triang(a, θ, b) distribution to a two sided power TSP(a, θ, b, n) distribution

$$f_X(x|a,\theta,b,n) = \frac{n}{b-a} \times \begin{cases} \left(\frac{x-a}{\theta-a}\right)^{n-1}, & a < x \leq \theta, \\ \left(\frac{b-x}{b-\theta}\right)^{n-1}, & \theta \leq x < b, \end{cases} \tag{5}$$

by the inclusion of an additional parameter $n > 0$ describing a power-law behavior in both tails. For $n = 2$ and $n = 1$ the distribution (5)

reduces to the $Triang(a, \theta, b)$ and Uniform$[a, b]$ distributions, respectively. The following expressions for the mean and the variance follow from (5):

$$E[X] = \frac{a + (n-1)\theta + b}{n+1}, Var[X] = \frac{n(b-a)^2 - 2(n-1)(b-\theta)(\theta - a)}{(n+2)(n+1)^2}. \tag{6}$$

Interestingly, one immediately observes that by substituting $n = s+1$ in (6), the beta mean value (4) and TSP mean value in (6) coincide. Moreover, recalling that $T \sim Beta(a, b, \alpha, \beta)$ given by (1) and $X \sim TSP(a, m, b, n)$ given by (5) and observing that for $s = 4$ or $n = 5$ the mean values in (4) and (6) agree and reduce to the PERT mean $E[T]$ in (2) as suggested by Malcolm et al. back in 1959, one might indeed conclude that they were lucky in this respect. However, observing that the variance in (4) for $s = 4$ is quite different from the PERT variance in (2), Malcolm et al. [1] were after all not so lucky. Moreover, after some algebraic manipulations using variances in (4) and (6) it follows that:

$$Var[T] - Var[X] = \frac{(n-1)(b-\theta)(\theta-a)}{(n+2)(n+1)} = \begin{cases} \leq 0, & 0 \leq n < 1, \\ > 0, & n > 1. \end{cases} \tag{7}$$

Hence, in the unimodal domains of the TSP distribution (5), $n > 1$, and the beta distributions (1), $s > 0$, with parameterization (3), the variance of the TSP distribution is strictly less than the PERT variance modification of Herrerías [9] given by (4). The result (7) is consistent with the TSP distributions being more "peaked" than the beta distribution (see, e.g. Kotz and Van Dorp [12]). Summarizing, given that an expert only provides lower bounds a and b and most likely value m, additional alternatives are provided in terms of the TSP(n) pdf's (5), $n \neq 2$, besides the existing beta and triangular pdf options, and one is left to wonder which one of these to use, perhaps extending the 50-year old controversy surrounding the use of (2).

The context of the controversy alluded to above deals with the larger domain of distribution selection and parameter elicitation via expert judgment, in particular those distributions with bounded support. In a recent survey paper, a leading Bayesian statistician O'Hagan [13] explicitly mentions a need for advances in elicitation techniques for prior distributions in Bayesian Analyses, but also acknowledges the importance of their development for those areas where the elicited distribution cannot be combined with evidence from data, because the expert opinion is essentially all the available knowledge. Garthwaite, Kadana and O'Hagan [14] provide a comprehensive review on the topic of eliciting probability distributions dealing with a wide variety of topics, such as e.g. the elicitation process, heuristics and biases, fitting distributions to an expert's summaries, expert calibration and group elicitation methods. Experts are, as a rule, classified into two, usually unrelated, groups: 1) substantive experts (also known as technical experts or domain experts) who are knowledgeable about the subject

matter at hand and 2) normative experts possessing knowledge of the appropriate quantitative analysis techniques (see, e.g., De Wispelare et al. [15] and Pulkkinen and Simola [16]). In the absence of data and in the context of decision/simulation and uncertainty analyses, substantive experts are used (often by necessity) to specify input distributions albeit directly or indirectly with the aid of a normative expert. The topic of this paper deals with fitting specific parametric distributions to a set of summaries elicited from an expert.

In Section 2, we provide an overview of indirect elicitation procedures for TSP PDF (5) parameters and their generalizations developed in Kotz and Van Dorp [17], Van Dorp et al. [18] and Herrerías et al. [19]. Firstly, we shall present an indirect elicitation procedure for the bound parameters a, b and tail parameter n of TSP PDF's (5). It has the specific advantage of not requiring bounds elicitation whom may not fall within the realm of expertise of a substantive expert. Next, we present the indirect elicitation of both tail parameter of a generalization of TSP distribution allowing for separate power law behavior in both tails. This procedure was presented in Herrerías et al. [19], but does require the bounds a and b to be available. We return to indirect bounds and power tail parameter elicitation for generalized trapezoidal uniform (GTU) distributions given lower and upper quantile estimates and a modal range specification. A substantive expert may be more comfortable with specifying a modal range rather than having to specify a single point estimate as required in (2), (3) and (5). The GTU elicitation procedure was developed in detail in Van Dorp et al. [18]. Finally, in Section 3, we shall demonstrate via an illustrative Monte Carlo analysis for the completion time of an 18 node activity network a potential difference between project completion times that could result when requiring experts to specify a single most likely estimate rather than allowing for a modal range specification.

2 PARAMETER ELICITATION ALGORITHMS FOR TSP DISTRIBUTIONS AND SOME GENERALIZATIONS

Let $X \sim TSP(\Theta)$ with PDF (5), where $\Theta = \{a, \theta, b, n\}$. The main advantage of the PDF (5) over the beta PDF (1) is that it has a closed form CDF expressible using only elementary functions:

$$F_X(x|\Theta) = \begin{cases} \frac{\theta-a}{b-a}\left(\frac{x-a}{\theta-a}\right)^n, & \text{for } a < x < \theta, \\ 1 - \frac{b-\theta}{b-a}\left(\frac{b-x}{b-\theta}\right)^n, & \text{for } \theta \leq x < b. \end{cases} \tag{8}$$

Suppose a lower and upper percentiles a_p, b_r and most likely value θ for X are pre-specified in a manner such that $a_p < \theta < b_r$. Kotz and Van

Dorp [17] showed that a unique bounds a and b solution

$$a \equiv a\{q(n)\} = \frac{a_p - \theta \sqrt[n]{p/q(n)}}{1 - \sqrt[n]{p/q(n)}},\ b \equiv b\{q(n)\} = \frac{b_r - \theta \sqrt[n]{\frac{1-r}{1-q(n)}}}{1 - \sqrt[n]{\frac{1-r}{1-q(n)}}} \tag{9}$$

exists, given a value for the parameter $n > 0$, where $q(n) = \Pr(X < \theta)$. Herein we shall use the notation $\sqrt[n]{x} = x^{1/n}$ even when $n > 0$ is non-integer valued. The unique value for $\Pr(X < \theta)$ follows by solving for $q(n)$ from the equation

$$q(n) = \frac{(\theta - a_p)\Big(1 - \sqrt[n]{\frac{1-r}{1-q(n)}}\Big)}{(b_r - \theta)\Big(1 - \sqrt[n]{\frac{p}{q(n)}}\Big) + (\theta - a_p)\Big(1 - \sqrt[n]{\frac{1-r}{1-q(n)}}\Big)}, \tag{10}$$

using a bisection method with starting interval $[p, r]$. When $n \downarrow 0$,

$$q(n) \to q(0) = (\theta - a_p)/(b_r - a_p) \tag{11}$$

and when $n \to \infty$, $q(n)$ converges to the unique solution $q(\infty)$ of the equation

$$\frac{q(\infty)}{q(0)} \log\Big\{\frac{q(\infty)}{p}\Big\} = \frac{1 - q(\infty)}{1 - q(0)} \log\Big\{\frac{1 - q(\infty)}{1 - r}\Big\}. \tag{12}$$

This equation, similar to $q(n)$ in (10), may be solved for using a bisection method with starting interval $[p, r]$. The PDF (5) itself, satisfying $a_p < \theta < b_r$, converges to a Bernoulli distribution with point mass $q(0)$ at a_p when $n \downarrow 0$ and when $n \to \infty$ converges to an asymmetric Laplace distribution

$$f_X(x|a_p, \theta, b_r) = \begin{cases} q(\infty)\mathcal{A}Exp\Big\{-\mathcal{A}(\theta - x)\Big\}, & x \le \theta, \\ \{1 - q(\infty)\}\mathcal{B}Exp\Big\{-\mathcal{B}(x - \theta)\Big\}, & x > \theta, \end{cases} \tag{13}$$

where the coefficients $\mathcal{A}$ and $\mathcal{B}$ are

$$\mathcal{A} = \frac{\log\big\{\frac{q(\infty)}{p}\big\}}{\theta - a_p} \text{ and } \mathcal{B} = \frac{\log\big\{\frac{1-q(\infty)}{1-r}\big\}}{b_r - \theta}. \tag{14}$$

See also Kotz and Van Dorp [20].

Summarizing, the information $a_p < \theta < b_r$ does not uniquely specify a member within the TSP family. Kotz and Van Dorp [17] suggest the elicitation of an additional quantile $a_p < x_s < b_r$ to indirectly elicit the remaining parameter n. They solve for a, b and n via an eight step algorithm. Its details are provided in Kotz and Van Dorp [17] and a software implementation of this algorithm is available from the author upon request. Setting $a_{0.10} = 6.5$, $x_{0.80} = 10\frac{1}{4}$, $b_{0.90} = 11\frac{1}{2}$ and $\theta = 7$ we have:

$$n \approx 3.873,\ q(n) = 0.209,\ a\{q(n)|n\} \approx 4.120, b\{q(n)|n\} \approx 17.878. \tag{15}$$

Figure 1 displays the TSP distribution with most likely value $\theta = 7$ and parameter values (15).

2.1 GTSP parameter elicitation algorithm

Kotz and Van Dorp [12] briefly mentioned generalized $GTSP(\Theta)$ distributions with PDF

$$f_X(x|\Theta) = \mathcal{C}(\Theta) \times \begin{cases} \left(\frac{x-a}{\theta-a}\right)^{m-1}, & \text{for } a < x < \theta, \\ \left(\frac{b-x}{b-\theta}\right)^{n-1}, & \text{for } \theta \leq x < b, \end{cases} \tag{16}$$

where $\Theta = \{a, \theta, b, m, n\}$ and

$$\mathcal{C}(\Theta) = \frac{mn}{(\theta-a)n + (b-\theta)m}. \tag{17}$$

They reduce to $TSP(\Theta)$ PDF's (5) when $m = n$ and were studied in more detail by Herrerías et al. [19]. Their CDF's follow from (16) as:

$$F_X(x|\Theta) = \begin{cases} \pi(\Theta)\left(\frac{x-a}{\theta-a}\right)^{m}, & \text{for } a < x < \theta, \\ 1 - [1-\pi(\Theta)]\left(\frac{b-x}{b-\theta}\right)^{n}, & \text{for } \theta \leq x < b. \end{cases} \tag{18}$$

where

$$\pi(\Theta) = (\theta - a)\mathcal{C}(\Theta)/m. \tag{19}$$

To indirectly elicit the power parameters m and n, Herrerias et al. [19] also suggest eliciting a lower quantile $a_p < \theta$ and an upper quantile $b_r > \theta$. Similar to the PERT mean and variance (2), however, lower and upper bounds a, b and a most likely estimate θ must have been directly pre-elicited. The parameters m and n are next solved from the following set of non-linear equations (the quantile constraints):

$$\begin{cases} F(a_p|\Theta) = \pi(\Theta)\left(\frac{a_p-a}{\theta-a}\right)^{m} = p, \\ F(b_r|\Theta) = 1 - [1-\pi(\Theta)]\left(\frac{b-b_r}{b-\theta}\right)^{n} = r. \end{cases} \tag{20}$$

Herrerias et al. [19] showed that the first (second) equation in (20) has a unique solution $m^{\bullet}$ for every fixed value of $n > 0$ and thus it defines an implicit continuous function $\xi(n)$ such that the parameter combination $\{\theta, m^{\bullet} = \xi(n),\, n\}$ satisfies the first quantile constraint for all $n > 0$. This unique solution $m^{\bullet}$ may be solved for by employing a standard root finding algorithm such as, e.g., the Newton-Raphson method (Press et al. [21]) or a commercially available one such as, e.g., GoalSeek in Microsoft Excel. Analogously, the second equation defines an implicit continuous function $\zeta(m)$ such that the parameter combination $(\theta, m,\, n^{\bullet} = \zeta(m))$ satisfies the second quantile constraint for all $m > 0$. By successively solving for the lower and upper quantile constraint given a value for n or m, respectively, an algorithm can be formulated that solves (20). Details are provided in Herrerias et al. [19]. Setting $a = 2$, $\theta = 7$, $b = 15$, $a_{0.10} = 4\frac{1}{4}$, and $b_{0.90} = 11$ in (20) yields the power parameters

$$m \approx 1.883 \text{ and } n \approx 2.460. \tag{21}$$

Figure 1 displays the GTSP distribution with lower and upper bounds $a = 2$ and $b = 15$, most likely value θ and the power parameter values (21).

2.2 GTU parameter elicitation procedure

Van Dorp et al. [18] considered Generalized Trapezoidal Uniform (GTU) distributions. Letting $X \sim GTU(\Theta)$, where $\Theta = \{a, \theta_1, \theta_2, b, m, n\}$, they have for its pdf:

$$f_X(x|\Theta) = \mathcal{C}(\Theta) \times \begin{cases} \left(\frac{x-a}{\theta_1 - a}\right)^{m-1}, & \text{for } a \leq x < \theta_1, \\ 1, & \text{for } \theta_1 \leq x < \theta_2, \\ \left(\frac{b-x}{b-\theta_2}\right)^{n-1}, & \text{for } \theta_2 \leq x < b, \end{cases} \tag{22}$$

where the normalizing constant $\mathcal{C}(\Theta)$ is given by

$$\mathcal{C}(\Theta) = \frac{mn}{(\theta_1 - a)n + (\theta_2 - \theta_1)mn + (b - \theta_2)m}. \tag{23}$$

Defining stage probabilities $\pi_1 = \Pr(X \leq \theta_1)$, $\pi_2 = \Pr(\theta_1 < X \leq \theta_2)$, and $\pi_3 = \Pr(X > \theta_1)$, one obtains from (22) and (23):

$$\begin{cases} \pi_1(\Theta) = \mathcal{C}(\Theta)(\theta_1 - a)/m, \\ \pi_2(\Theta) = \mathcal{C}(\Theta)(\theta_2 - \theta_1), \\ \pi_3(\Theta) = \mathcal{C}(\Theta)(b - \theta_2)/n. \end{cases} \tag{24}$$

Utilizing the stage probabilities $\pi_i(\Theta)$, $i = 1, 2, 3$, one obtains the following convenient form for the CDF of (22)

$$F_X(x|\Theta) = \begin{cases} \pi_1(\Theta)\left(\frac{x-a}{\theta_1 - a}\right)^m, & a \leq x \leq \theta_1, \\ \pi_1(\Theta) + \pi_2(\Theta)\frac{x-\theta_1}{\theta_2 - \theta_1}, & \theta_1 < x \leq \theta_2 \\ 1 - \pi_3(\Theta)\left(\frac{b-x}{b-\theta_2}\right)^n, & \theta_2 < x \leq b, \end{cases} \tag{25}$$

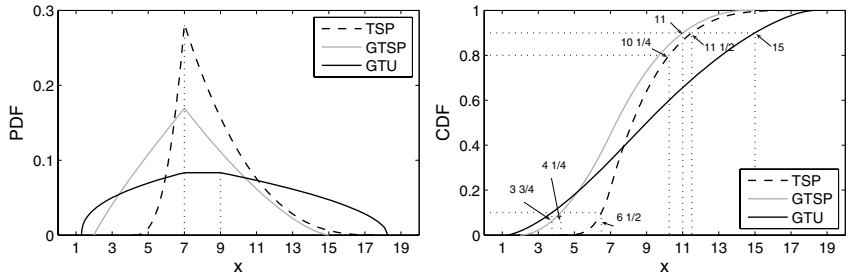
and for its quantile function

$$F_X^{-1}(y|\Theta) = \begin{cases} a + (b-a)\sqrt[m]{\frac{y}{\pi_1(\Theta)}}, & 0 \leq y \leq \pi_1(\Theta), \\ b + (c-b)\frac{y - \pi_1(\Theta)}{\pi_2(\Theta)}, & \pi_1(\Theta) < y \leq 1 - \pi_3(\Theta), \\ d - (d-c)\sqrt[n]{\frac{1-y}{\pi_3(\Theta)}}, & 1 - \pi_3(\Theta) < y \leq 1. \end{cases} \tag{26}$$

The $GTU(\Theta)$ distributions reduce to trapezoidal distributions studied by Pouliquen [22] by setting $m = n = 2$ to $GTSP(\Theta)$ distributions given by (16) and (17) by setting $\theta_1 = \theta_2$, and to $TSP(\Theta)$ distributions given by (5) by setting $\theta_1 = \theta_2 = \theta$ and $m = n$ in (22) and (23).

It shall be assumed here that the lower and upper bound parameters a and b and tail parameters m and n are unknown and that they need to be determined from (i) a directly elicited modal range $[\theta_1, \theta_2]$, (ii) the relative likelihoods π_2/π_1 and π_2/π_3 (or their reciprocals), and (iii) a lower $a_p < \theta_1$ and upper $b_r > \theta_2$ quantiles. The first (second) relative likelihood may be elicited by asking how much more likely it is for X to be within its modal

range $[\theta_1, \theta_2]$ than being less (larger) than it. Stage probabilities (24) π_i, $i = 1, 2, 3$, next follow with the restriction they must sum to 1. This manner of elicitating of π_i for $i = 1, 2, 3$ is analogous to the fixed interval elicitation method mentioned in Garthwaite, Kadane and O'Hagan [14].



(a) (b)

FIGURE 1. PDF's (a) and CDF's (b) of TSP, GTSP and GTU distributions with parameter settings (15), (21) and (29)

Van Dorp et al. [18] showed that a unique solution for the power parameters m and n may be obtained from the equations $a^*(m) = \tilde{a}(m)$ and $b^*(n) = \tilde{b}(n)$, respectively, where

$$\begin{cases} a^*(m) \equiv \theta_1 - m\frac{\pi_1}{\pi_2}(\theta_2 - \theta_1), & \tilde{a}(m) \equiv a_p - \frac{\sqrt[m]{p/\pi_1}}{1-\sqrt[m]{p/\pi_1}}(\theta_1 - a_p), \\ b^*(n) \equiv \theta_2 + n\frac{\pi_3}{\pi_2}(\theta_2 - \theta_1), & \tilde{b}(n) \equiv b_r + \frac{\sqrt[n]{(1-r)/\pi_3}}{1-\sqrt[n]{(1-r)/\pi_3}}(b_r - \theta_2), \end{cases} \tag{27}$$

π_i, with $i = 1$, 2, and 3, are given by (25), and provided

$$\begin{cases} a_p > b - \xi(c - b), & \text{where } \xi = \frac{\pi_1}{\pi_2}\log\left(\frac{\pi_1}{p}\right) > 0, \\ d_r < c + \psi(c - b), & \text{where } \psi = \frac{\pi_3}{\pi_2}\log\left(\frac{\pi_3}{1-r}\right) > 0. \end{cases} \tag{28}$$

The equations $a^*(m) = \tilde{a}(m)$ and $b^*(n) = \tilde{b}(n)$ may be solved for using a standard root finding algorithm such as, e.g., the Newton-Raphson method (Press et al. [21]) or a commercially available one such as, e.g., GoalSeek in Microsoft Excel. No solution for power parameters m and n exist when conditions in (28) are not met. After solving for m, the lower bound a follows by substitution of m in $a^*(m)$ or $\tilde{a}(m)$. Solving for the upperbound b is analogous, but utilizes the expressions for $b^*(n)$ or $\tilde{b}(n)$. Setting $[\theta_1, \theta_2] = [7, 9]$, $\pi_2/\pi_1 = 1/2$, $\pi_2/\pi_3 = 1/3$, $a_{0.10} = 3\frac{3}{4}$ and $b_{0.90} = 15$ in (27) yields the tail and lower and upper bound parameters

$$m \approx 1.423,\ n \approx 1.546,\ a \approx 1.306 \text{ and } b \approx 18.273. \tag{29}$$

Figure 1 displays the GTU distribution with modal range $[\theta_1, \theta_2] = [7, 9]$ and parameter values (29). Please observe in Figure 1 that both TSP and

GTSP distributions posses mode $\theta = 7$, whereas the GTU distribution has a modal range [7,9]. Quantile values for the TSP, GTSP and GTU examples in this section are indicated in Figure 1b. Elicited modal (quantile) values are indicated in Figure 1a (Figure 1b).

3 AN ILLUSTRATIVE ACTIVITY NETWORK EXAMPLE

We shall demonstrate via an illustrative Monte Carlo analysis for the completion time of an 18 node activity network from Taggart [23], depicted in Figure 2, a potential difference between project completion times that could result when requiring experts to specify a single most likely estimate rather than allowing for a modal range specification. We shall assume that lower and upper quantiles $a_{0.10}$ and $b_{0.90}$ in Table 1 have been elicited via an expert judgment for each activity in the project network. We shall investigate four scenarios of mode specification for the activity durations in the project network, keeping their lower and upper quantiles $a_{0.10}$ and $b_{0.90}$ fixed. In the first scenario "GTU" activity duration uncertainty is modeled using a GTU distribution. The modal range $[\theta_1, \theta_2]$ is specified in Table 1. For all activities, a relative likelihood of 2.75 (1.25) is specified for the right tail (left tail) as compared to the modal range $[\theta_1, \theta_2]$. From the relative likelihoods it immediately follows that the lower bounds θ_1 of the modal ranges in Table 1 equal the first quartile (probability $\frac{1}{4}$) of the activities, whereas a $\frac{1}{5}$ probability is specified throughout for the modal range $[\theta_1, \theta_2]$. Hence, the upper bounds θ_2 of the modal ranges are the 45-th percentiles of the activity durations and thus are strictly less than their median values. Moreover, all activity durations are right skewed (having a longer tail towards the right). We solve for the lower and upper bounds a and b using the procedure described in Section 2.2.

The next three scenarios involve limiting cases when activity duration uncertainties are distributed as a two-sided power (TSP) distribution with the PDF (5). Recall from Section 2 that Kotz and Van Dorp [17] have shown that for every $n > 1$ in (5), a unique unimodal TSP distribution can be fitted given a lower quantile $a_{0.10}$, an upper quantile $b_{0.90}$ and a most likely value θ such that $a_{0.10} < \theta < b_{0.90}$. For $n \downarrow 1$, the fitted TSP distribution reduces to a uniform distribution with the bounds

$$a = \frac{0.90a_{0.10} - 0.10b_{0.90}}{0.80} \text{ and } b = \frac{0.90b_{0.90} - 0.10a_{0.10}}{0.80}. \qquad (30)$$

We shall use bounds (30) for the second scenario designated "Uniform" combined with the values for $a_{0.10}$ and $b_{0.90}$ in Table 1. The uniform distribution with bounds (29) actually has the smallest variance amongst pdf's (5) given the constraint set by $a_{0.10} < \theta < b_{0.90}$ and their fixed values.

For $n \to \infty$ and with specified values $a_{0.10} < \theta < b_{0.90}$, the TSP distribution (5) converges to an asymmetric Laplace distribution (13) with parameters $a_{0.10}, \theta, b_{0.90}$ and $q(\infty)$, where $q(\infty)$ is the limiting probability of being less than the mode θ and the unique solution to Equation (12).

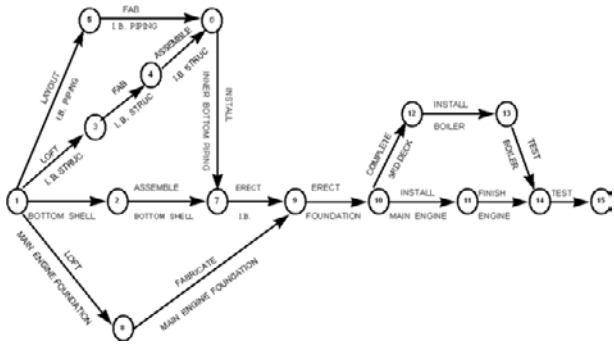


FIGURE 2. Example project network from Taggart [23]

This asymmetric Laplace distribution has the largest variance amongst the TSP distributions (5) given the constraint $a_{0.10} < \theta < d_{0.90}$ and their preset values. Hence, for our third scenario "Laplace 1" we set $\theta = \theta_1$, specified in Table 1, and use the values $a_{0.10}$ and $b_{0.90}$ in Table 1 to determine the remaining parameter $q(\infty)$. Similarly, we obtain the fourth scenario "Laplace 2" by setting $\theta = \theta_2$. Note that our first two scenarios "GTU" and "Uniform" are consistent with the mode specifications $a_{0.10} < \theta < b_{0.90}$ in the third and fourth scenarios "Laplace 1" and "Laplace 2", respectively. That is, in all the scenarios the activity durations have the lower and upper quantiles $a_{0.10}$ and $b_{0.90}$ in common and a mode at $\theta = \theta_1$ ($\theta = \theta_2$) for the third (fourth) scenario.

Now we shall generate the CDF of the completion time distribution of the project presented in Figure 2 for each of these scenarios "GTU", "Uniform", "Laplace 1" and "Laplace 2" by employing the Monte Carlo technique (Vose [24]) involving 25,000 independent samples from the activity durations and subsequently applying the critical path method (CPM) (see e.g. Winston [4]). To avoid the occurence of negative activity durations in the sampling routine as a result of the infinite support of the Laplace distributions, a negative sampled activity duration is set to be equal to zero. Consequently, for each scenario we obtain an output sample of size 25000 for the completion time of the project network in Figure 2 from which one can empirically estimates its completion time distribution. The resulting CDF's for the four scenarios described above are depicted in Figure 3. Among the scenario's in Figure 3 only the scenario "Uniform" has symmetric activity duration distributions. The activity durations of all other scenarios are all right skewed with a mean value less than that of the same activity in the "Uniform" scenario. This explains why the completion time distribution of the "Uniform" scenario is located substantially to the right of all the other scenarios. Moreover, as explained above, the variances of activity durations

Activity name	$a_{0.10}$	θ_1	θ_2	$b_{0.90}$
Shell: loft	22	25	28	41
Shell: Assemble	35	38	41	54
I.B.Piping: Layout	22	25	28	41
I.B.Piping: Fab.	6	8	10	19
I.B.Structure: Layout	22	25	28	41
I.B.Structure:Fab.	16	18	20	29
I.B.Structure:Assemb.	11	13	15	24
I.B.Structure:Install	6	8	10	19
Mach Fdn. Loft	26	29	32	45
Mach Fdn. Fabricate	31	34	37	50
Erect I.B.	28	31	34	47
Erect Foundation	6	8	10	19
Complete 3rd DK	4	6	8	17
Boiler: Install	7	9	11	20
Boiler: Test	9	11	13	22
Engine: Install	6	8	10	19
Engine: Finish	18	21	24	37
Final Test	14	17	20	33

TABLE 1. Data for modeling the uncertainty in activity durations for the project network presented in Figure 2

in the "Uniform" scenario are smaller than those of the activities in the other one. Thus it explains why its project completion time CDF is the steepest.

The largest discrepancy between the CDF's in Figure 3 occurs between the "Uniform" and "Laplace 1" and equals ≈ 0.24 observed at ≈ 194 days. Hence, certainly the specification of lower and upper quantiles $a_{0.10}$ and $b_{0.90}$ and a most likely value θ seems to be insufficient to determine a PDF in the family (5). Note that the project completion time CDF of the "GTU" scenario in Figure 3 for the most part is sandwiched between those of the "Laplace 1" and "Laplace 2" scenarios with a maximal difference of ≈ 0.04 (≈ 0.07) between its CDF and the "Laplace 1" ("Laplace 2") CDF's observed at approximately 187 days (197 days).

Finally, note that in Figure 3 the project completion time of 149 days following from the CPM using only the most likely values of θ_1 in Table 1, is represented by the bold vertical dashed line "CPM 1". Similarly, a completion time of 171 days follows using only the most likely values of θ_2 in Table 1 is indicated by the bold "CPM 2" line. Since the values of θ_1 are less than the median for all 18 activities in Table 1 (in addition to having right skewness), we observe from Figure 3 that the probability of achieving the "CPM 1" completion time of 149 days is negligible. For the "CPM 2" completion time of 171 days these probabilities are less than $\approx 10\%$ for all four scenarios. Although the skewness of the activity distributions in Table 1

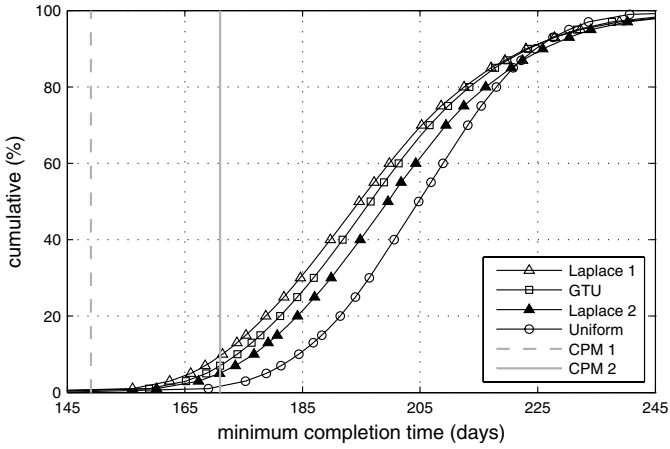


FIGURE 3. Comparison of CDF's of the completion times for the project in Figure 2

may perhaps be somewhat inflated, a case could definitely be made that a skewness towards the lower bound may appear in assessed activity time distributions in view of a potential motivational bias of the substantive expert. These CPM results further reinforce the observation that in applications uncertainty results ought to be communicated to decision makers.

4 CONCLUDING REMARKS

A discussion some 50 years ago about the appropriateness of using the PERT mean and variance (2) utilizing either beta or triangular pdfs, was followed by a concern by others some 20 years later or more (e.g. Selvidge [25] and Keefer and Verdini [26]) regarding the elicitation of lower and upper bounds a, b of a bounded uncertain phenomenon, since these typically do not fall within the realm of experience of an substantive expert. When instead eliciting a lower and upper quantiles a_p and b_r and a most likely value θ, however, even within the two-sided power (TSP) family of distribution with bounded support, infinitely many options exist that match these constraints. Hence, one arrives at the conclusion that additional information needs to elicited from the substantive expert for further uncertainty distribution specification. In case of the TSP family of distributions, Kotz and Van Dorp [17] suggested the elicitation of an additional quantile to uniquely identify its lower and upper bounds a and b and power parameter n. Even when relaxing the TSP PDF or PERT requirement of specifying a single mode θ to allow for a modal range specification $[\theta_1, \theta_2]$ of a generalized trapezoidal uniform (GTU) distributions, a lower quantile $a_p < \theta_1$ and upper quantile $b_r > \theta_2$ specification is not a sufficient information to

determine its lower and upper bounds $a < a_p$ and $b > b_r$ and its power parameters m and $n > 0$. Van Dorp et al. [18] suggest to elicit in addition two relative likelihoods regarding the three stages of the GTU distribution to solve for these parameters.

Summarizing, lower and upper bounds specification or lower and upper quantiles specification combined with providing a single modal value, or even a modal range, does not uniquely determine an uncertainty distribution. In my opinion, this lack of specificity is one of the root causes regarding the controversy alluded to in the introduction of this paper surrounding the continued use of the PERT mean and variance (2) or other common arguments amongst practitioners regarding whether to use beta, triangular (or TSP) distributions to describe a bounded uncertain phenomena.

Acknowledgments

I am indebted to Samuel Kotz who has been gracious in donating his time to provide comments and suggestions in the development of various sections presented in this paper. I am also thankful to the referee and the editor whose comments and editorial support improved the presentation of an earlier version considerably.

Bibliography

[1] D. G. Malcolm, C. E. Roseboom, C. E. Clark, and W. Fazar. Application of a technique for research and development program evaluation. *Operations Research*, 7:646–649, 1959.

[2] T. Simpson. A letter to the Right Honourable George Earls of Maclesfield. President of the Royal Society, on the advantage of taking the mean of a number of observations in practical astronomy. *Philosophical Transactions*, 49(1):82–93, 1755.

[3] T. Simpson. An attempt to show the advantage arising by taking the mean of a number of observations in practical astronomy. *Miscellaneous Tracts on some curious and very interesting Subjects in Mechanics, Physical Astronomy and Speculative Mathematics*, pages 64–75, 1757.

[4] W. L. Winston. *Operations Research, Applications and Algorithms*. Duxbury Press, Pacific Grove, CA, 1993.

[5] J. Kamburowski. New validations of PERT times. *Omega, International Journal of Management Science*, 25(3):323–328, 1997.

[6] C. E. Clark. The PERT model for the distribution of an activity. *Operations Research*, 10:405–406, 1962.

[7] F. E. Grubbs. Attempts to validate certain PERT statistics or a 'picking on PERT'. *Operations Research*, 10:912–915, 1962.

[8] J. J. Moder and E. G. Rodgers. Judgment estimate of the moments of PERT type distributions. *Management Science*, 15(2):B76–B83, 1968.

[9] R. Herrerías. Utilización de Modelos Probabilísticos Alternativas para el Métedo PERT. Applicación al Análisis de Inversiones. *Estudios de Economía Aplicada*, pages 89–112, 1989.

[10] R. Herrerías, J. García, and S. Cruz. A note on the reasonableness of PERT

hypotheses. *Operations Research Letters*, 31:60–62, 2003.

[11] J. R. Van Dorp and S. Kotz. A novel extension of the triangular distribution and its parameter estimation. *The Statistician*, 51(1):63–79, 2002.

[12] S. Kotz and J. R. Van Dorp. *Beyond Beta, Other Continuous Families of Distributions with Bounded Support and Applications*. World Scientific Press, Singapore, 2004.

[13] A. O'Hagan. Research in elicitation. In S. K. Upadhyay, U. Singh, and D. K. Dey, editors, *In Bayesian Statistics and its Applications*, pages 375–382. Anamaya Publishers, New Delhi, 2006.

[14] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–700, 2005.

[15] A. DeWispelare, L. Herren, and R. T. Clemen. The use of probability elicitation in the high-level nuclear waste recognition program. *International Journal of Forecasting*, 11(1):5–24, 1995.

[16] U. Pulkkinen and K. Simola. An expert panel approach to support risk-informed decision making. Technical Report STUK-YTO-TR 129, Sateiluturvakeskus (Radiation and Nuclear Safety Authority of Finland STUK), Helsinki, Finland, 2009.

[17] S. Kotz and J. R. Van Dorp. A novel method for fitting unimodal continuous distributions on a bounded domain. *IIE Transactions*, 38:421–436, 2006.

[18] J. R. Van Dorp, S. Cruz, J. García, and R. Herrerías. An elicitation procedure for the generalized trapezoidal distribution with a uniform central stage. *Decision Analysis*, 4:156–166, 2007.

[19] J. M. Herrerías, R. Herrerías, and J. R. Van Dorp. The generalized two-sided power distribution. *Journal of Applied Statistics*, 35(5):573–587, 2009.

[20] S. Kotz and J. R. Van Dorp. A link between two-sided power and asymmetric laplace distributions: with applications to mean and variance approximations. *Statistics and Probability Letters*, 71:382–394, 2005.

[21] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering. *Numerical Recipes in Pascal*. Cambridge University Press, Cambridge, UK, 1989.

[22] L. Y. Pouliquen. *Risk analysis in project appraisal. World Bank Staff Occasional Papers*. Hopkins University Press, Baltimore, MD, 1970.

[23] R. Taggart. *Ship Design and Construction*. The Society of Naval Architects and Marine Engineers (SNAME), New York, 1980.

[24] D. Vose. *Quantitative Risk Analysis, A Guide to Monte Carlo Simulation Modeling*. Wiley, New York, 1980.

[25] J. E. Selvidge. Assessing the extremes of probability distributions by the fractile method. *Decision Sciences*, 11:493–502, 1980.

[26] D. L. Keefer and A. V. Verdini. Better estimation of PERT activity time parameters. *Management Science*, 39(9):1086–1091, 1993.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Finding proper non-informative priors for regression coefficients

H.R.N. van Erp* and P.H.A J.M. van Gelder
– Delft University of Technology, Delft, The Netherlands

Abstract. By using informational consistency requirements, Jaynes (1968) derives the form of maximal non-informative priors for regression coefficients, to be uniform. However, this result does not tell us what the limits of this uniform distribution should be. If we are faced with a problem of model selection this information is an integral part of the evidence, which is used to rank the various competing models. In this paper, we give some guidelines for choosing a parsimoneous proper uniform prior. It turns out that in order to construct such a parsimoneous prior one only needs to assign a maximal length to the dependent variable and minimal lengths to the independent variables, together with their maximal correlations.

1 INTRODUCTION

It is a known fact that in problems of Bayesian model selection improper priors may lead to biased conclusions. In this paper we first give a short introduction to the procedure of Bayesian model selection. We then demonstrate for a simple model selection problem, involving two regression models, how improper uniform priors for the regression coefficients will exclude automatically the model with the most regression coefficients. Having established the problematic nature of improper priors for this particular case we proceed to derive a parsimoneous proper uniform prior for univariate regression models, firstly, and then generalize this result to multivariate regression models, secondly.

2 BAYESIAN MODEL SELECTION

We will give here a simple outline of the procedure of Bayesian model selection. Let $p\left(\theta|I\right)$ be the prior of some parameter θ conditional on the background information I. Let $p\left(D|\theta,M,I\right)$ be the probability of the data D conditional on the value of parameter θ, the particular model M used, and the background information I; the probability of the data is also known

*corresponding author: Structural Hydraulic Engineering and Probabilistic Design, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5, 2600 Delft, the Netherlands; telephone: +31-(0)15 27 89448, e-mail: h.r.n.vanerp@tudelft.nl

as the likelihood of the parameter θ. Let $p(\theta|D, M, I)$ be the posterior distribution of the parameter θ conditional on the data D, the particular model M used, and the background information I. We then have that

$$p(\theta|D, M, I) = \frac{p(\theta|I)\, p(D|\theta, M, I)}{\int p(\theta|I)\, p(D|\theta, M, I)\, d\theta} = \frac{p(\theta|I)\, p(D|\theta, M, I)}{p(D|M, I)} \quad (1)$$

where

$$p(D|M, I) \equiv \int p(\theta|I)\, p(D|\theta, M, I)\, d\theta \quad (2)$$

is the marginalized likelihood of the model M, also known as the evidence of model M.

Say we have m different models, $M_1, \ldots, M_m$. Then we may compute m different evidence values, $p(D|M_j, I)$ for $j = 1, \ldots, m$. Let $p(M_j|I)$ be the prior of model M_j conditional on the background information I. Let $p(M_j|D, I)$ be the posterior distribution of the model M_j conditional on the data D and the background information I. We then have that

$$p(M_j|D, I) = \frac{p(M_j|I)\, p(D|M_j, I)}{\sum p(M_j|I)\, p(D|M_j, I)}. \quad (3)$$

Note that if $p(M_j|I) = p(M_k|I)$ for $j \neq k$, we have that (3) reduces to

$$p(M_j|D, I) = \frac{p(D|M_j, I)}{\sum p(D|M_j, I)}. \quad (4)$$

Stated differently, if we assign equal prior probabilities to our different models, the posterior probabilities of these models reduce to their normalized evidence values, that is, the models may be ranked by their respective evidence values [1].

3 THE PROBLEM OF IMPROPER PRIORS IN MODEL SELECTION

There is a long tradition of the use of improper uniform priors for regression coefficients, that is, location parameters, in problems of parameter estimation [2, 3, 4]. However, in problems of model comparison between competing regression models one generally must take care not to use improper priors, be they uniform or not, because this may introduce inverse infinities in the evidence factors which may not cancel out if one proceeds to compute the posterior probabilities of the respective models. We will demonstrate this fact and its consequences below with a simple example in which we assign variable uniform priors to the respective regression coefficients.

Suppose that we want to compare two regression models. Say,

$$M_a : \mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{e}_a, \quad M_b : \mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \mathbf{e}_b, \quad (5)$$

where $\mathbf{e}_a = (e_{a1}, \ldots, e_{aN})$, $\mathbf{e}_b = (e_{b1}, \ldots, e_{bN})$, and $e_{ai}\tilde{e}_{bi}\tilde{N}(0, \sigma)$ for $i = 1, \ldots, N$, for some known value of σ. Let the independent priors of β_1 and β_2 be given as

$$p(\beta_1|I) = p(\beta_2|I) = \frac{1}{2A}, -A \leq \beta_1, \beta_2 \leq A. \tag{6}$$

Let the likelihoods be given, respectively, as

$$p(\mathbf{y}|\beta_1, \sigma, M_a, I) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}_1\beta_1)^T(\mathbf{y} - \mathbf{x}_1\beta_1)\right] \tag{7a}$$

$$p(\mathbf{y}|\beta, \sigma, M_b, I) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)\right] \tag{7b}$$

where $X \equiv [\mathbf{x}_1\ \mathbf{x}_2]$ and $\beta \equiv [\beta_1\ \beta_2]^T$. Combining the priors (6) with the likelihoods (7), and integrating out the unknown β's, β_1 and β_2, we get the following two evidence values [4]:

$$p(\mathbf{y}|\sigma, M_a, I) = \frac{1}{2A} L_1 \tag{8a}$$

where $L_1 \equiv (2\pi\sigma^2)^{-(N-1)/2} \|\mathbf{x_1}\| \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{x}_1\hat{\beta}_1)^T(\mathbf{y} - \mathbf{x}_1\hat{\beta}_1)\right]$ and $\hat{\beta}_1 \equiv \mathbf{x}_1^T\mathbf{y}/\mathbf{x}_1^T\mathbf{x}_1$, and

$$p(\mathbf{y}|\sigma, M_b, I) = \frac{1}{(2A)^2} L_2 \tag{8b}$$

where $L_2 \equiv (2\pi\sigma^2)^{-(N-2)/2} \left|X^TX\right|^{1/2} \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y} - X\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - X\hat{\boldsymbol{\beta}}\right)\right]$, $\hat{\boldsymbol{\beta}} \equiv \left(X^TX\right)^{-1}X^T\mathbf{y}$, and $\left|X^TX\right|^{1/2}$ is the square of the determinant of the inner product of the matrix X.

Now, if we assign equal prior probabilities to models M_a and M_b we may subsitute the (8) into (4) and so get

$$p(M_a|\mathbf{y}, \sigma, I) = \frac{L_1}{L_1 + L_2/2A}, \tag{9a}$$

$$p(M_b|\mathbf{y}, \sigma, I) = \frac{L_2/2A}{L_1 + L_2/2A} \tag{9b}$$

Looking at (9) we see that assigning improper uniform priors, that is, letting $A \to \infty$ in (6), will make $p(M_a|\mathbf{y}, \sigma, I) \to 1$ and $p(M_b|\mathbf{y}, \sigma, I) \to 0$. In Bayesian model selection of competing regression models care should be taken not to take unnecessarily large values of A in (6), since this will penalize those regression models which carry the most regression coefficients. In a limit of infinity, as (6) becomes improper, the regression model which

has the least regression coefficients will always automatically be chosen over any model which has more regression coefficients.

Note that if we are comparing models (5) we may use an improper prior for the unknown parameter σ, since the inverse infinity introduced in doing this is shared by both models and, thus, will be cancelled out if we compute the posterior probabilities (9) of the respective regression models. Using an improper prior for σ we may easily integrate out this unknown parameter [4].

4 DERIVING PROPER UNIFORM PRIORS FOR THE UNIVARIATE CASE

We have seen above that overly large priors penalize models which carry more regression coefficients to the point of excluding them altogether in a limit where these priors become improper. In problems of Bayesian model selection parsimoneous proper priors for the regression coefficients should be used. In what follows we derive the, trivial, limits of the univariate uniform prior for a single regression coefficient. The extension to the multivariate case, which we will give in the next paragraph, is based upon the basic idea introduced here.

Say we wish to regress an dependent vector $\mathbf{y}$ upon an independent vector $\mathbf{x}$. Then, using matrix algebra, the regression coefficient β may be computed as:

$$\beta = \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}} = \frac{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}{\|\mathbf{x}\|^2} \cos\theta = \frac{\|\mathbf{y}\|}{\|\mathbf{x}\|} \cos\theta. \tag{10}$$

By examining (10) we see that β must lie in the interval

$$\frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}} (\cos\theta)_{\min} \leq \beta \leq \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}} (\cos\theta)_{\max}. \tag{11}$$

Since $(\cos\theta)_{\min} = -1$ and $(\cos\theta)_{\max} = 1$, interval (11) reduces to:

$$-\frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}} \leq \beta \leq \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}}. \tag{12}$$

So, knowing the prior minimal length of the predictor, $\|\mathbf{x}\|_{\min}$, and the prior maximal length of the outcome variable, $\|\mathbf{y}\|_{\max}$, we may set the limits to the possible values of β. It follows that the proper non-informative prior of β must be the univariate uniform distribution with limits as given in (12):

$$p(\beta|I) = \frac{\|\mathbf{x}\|_{\min}}{2\|\mathbf{y}\|_{\max}}, \quad -\frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}} \leq \beta \leq \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}\|_{\min}} \tag{13}$$

Note that the prior (13) is a specific member of a more general familly of uniform priors (6).

5 DERIVING PROPER UNIFORM PRIORS FOR THE MULTIVARIATE CASE

We now derive the limits of the multivariate uniform prior for k regression coefficients. The basic idea used here is a generalization of the very simple idea that was used to derive the limits for the univariate case. This generalization will involve a transition from univariate line pieces to multivariate ellipsoids.

Say we have k independent predictors $\mathbf{x}_1,\ldots,\mathbf{x}_k$, that is, $\mathbf{x}_i^T\mathbf{x}_j = 0$ for $i \neq j$. Then we have that

$$\beta_i = \frac{\mathbf{x}_i^T\mathbf{y}}{\mathbf{x}_i^T\mathbf{x}_i} = \frac{\|\mathbf{y}\|}{\|\mathbf{x}_i\|}\cos\theta_i, \quad -\frac{\pi}{2} \leq \theta_i \leq \frac{\pi}{2}. \tag{14}$$

Because of the independence of the k independent variables we have that if one of the angles $\theta_i = 0$, then $\theta_j = \pi/2$ for $j \neq i$. It follows that all the possible values of β_i must lie in an k-variate ellipsoid centered at the origin and with respective axes of

$$r_i = \frac{\|\mathbf{y}\|_{\max}}{\|\mathbf{x}_i\|_{\min}}. \tag{15}$$

If we substitute (15) in the identity for the volume of an k-variate ellipsoid

$$V = \pi\left(\frac{4}{3}\right)^{k-2}\prod_{i=1}^{k} r_i. \tag{16}$$

We find that

$$V = \pi\left(\frac{4}{3}\right)^{k-2}\frac{\|\mathbf{y}\|_{\max}^k}{\prod_{i=1}^k \|\mathbf{x}_i\|_{\min}}. \tag{17}$$

Let $X \equiv [\mathbf{x}_1 \ \cdots \ \mathbf{x}_k]$. Then for k independent variables $\mathbf{x}_i$ the product of the norms is equivalent to the square root of the determinant of X^TX, that is,

$$\prod_{i=1}^{k}\|\mathbf{x}_i\| = \left|X^TX\right|^{1/2}, \tag{18}$$

which is also the volume of the parallelepiped defined by the vectors $\mathbf{x}_1$, ..., $\mathbf{x}_k$. Now in the case the k predictors $\mathbf{x}_1$, ..., $\mathbf{x}_k$, are not independent, that is, $\mathbf{x}_i^T\mathbf{x}_j \neq 0$ for $i \neq j$, we can transform them to an orthogonal basis $\tilde{\mathbf{x}}_1$, ..., $\tilde{\mathbf{x}}_k$, and $\tilde{X} \equiv [\tilde{\mathbf{x}}_1 \ \cdots \ \tilde{\mathbf{x}}_k]$, using the Gram-Schmidt orthogonalization process [5]. Since the volume of the parallelepid is invariant under a change of basis we have

$$\left|\tilde{X}^T\tilde{X}\right|^{1/2} = \left|X^TX\right|^{1/2}. \tag{19}$$

Thus substituting (19) into (18) we get, for both independent and dependent predictors

$$V = \pi\left(\frac{4}{3}\right)^{k-2}\frac{\|\mathbf{y}\|_{\max}^k}{|X^TX|_{\min}^{1/2}}. \tag{20}$$

Now, if we wish to assign a proper uniform prior to the regression coefficients $\beta_1, \ldots, \beta_k$, we may use the inverse of (20), that is,

$$p\left(\beta_1, \ldots, \beta_k | I\right) = \frac{1}{\pi}\left(\frac{3}{4}\right)^{k-2} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{k}}, \quad \beta_1, \ldots, \beta_k \in \text{Ellipsoid}, \tag{21}$$

where

$$\left|X^T X\right|_{\min}^{1/2} = \begin{vmatrix} 1 & (\cos\phi_{12})_{\max} & \cdots & (\cos\phi_{1k})_{\max} \\ (\cos\phi_{12})_{\max} & 1 & \cdots & (\cos\phi_{2k})_{\max} \\ \vdots & \vdots & \ddots & \vdots \\ (\cos\phi_{1k})_{\max} & (\cos\phi_{2k})_{\max} & \cdots & 1 \end{vmatrix}^{1/2} \prod_{i=1}^{k} \|\mathbf{x}_i\|_{\min}, \tag{22}$$

where $\cos\phi_{ij}$ is the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$. Looking at (21) and (22), we see that maximizing the area of our prior hypothesis is accomplished by maximizing the length of the dependent variable $\mathbf{y}$ and minimizing the determinant of the inner product of the matrix X, where the latter is accomplished by minimizing the lengths of the dependent variables $\mathbf{x}_1,\ldots,\mathbf{x}_k$ and maximizing the correlations $\cos\phi_{12}, \ldots, \cos\phi_{k-1,k}$ between the dependent variables.

6 DISCUSSION

By using informational consistency requirements Jaynes [3] derives the form of maximal non-informative priors for location parameters, that is, regression coefficients, to be uniform. However, this result does not tell us what the limits of this this uniform distribution should be, that is, what particular uniform distribution to use. Now, if we are just faced with a parameter estimation problem these limits of the non-informative uniform prior are irrelevant, since we may scale the product of the improper uniform prior and the likelihood to one, thus obtaining a properly normalized posterior. However, if we are faced with a problem of model selection the value of the uniform prior is an integral part of the evidence, which is used to rank the various competing models. We have given here some guidelines for choosing a parsimoneous proper uniform prior. It has turned out that in order to assign this prior one only needs to assign a maximal length to the dependent variable $\mathbf{y}$ and minimal lengths to the independent variables $\mathbf{x}_1, \ldots, \mathbf{x}_k$, together with their maximal correlations $\cos\phi_{12}, \ldots, \cos\phi_{k-1,k}$.

Bibliography

[1] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, 2003.

[2] H. J Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, third edition, 1961.

[3] E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, SSC-4(3):227–241, 1968.

[4] Arnold Zellner. *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, Chichester, 1971.

[5] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, Massachussets, 1993.

APPENDIX: Bayesian model selection, maximum likelihood selection, and Occam factors

Having derived a suitable parsimoneous proper non-informative uniform prior for the multivariate case, we now will take a closer look at the evidence values which result from using this prior. We will also discuss the connection between Bayesian model comparison and classical maximum likelihood model selection. To this end we will introduce the concept of the Occam factor.

Suppose we wish the compute the evidence of a specific model M, with

$$M : \mathbf{y} = X\beta + \mathbf{e}, \tag{23}$$

where $\mathbf{e} = (e_1, \ldots, e_N)$ and $e_i \tilde{N}(0, \sigma)$ for $i = 1, \ldots, N$, and for some known value of σ, then the corresponding likelihood (7b) is

$$p(\mathbf{y}|X, \beta, \sigma, M) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{(\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta)}{2\sigma^2}\right]. \tag{24}$$

Combining the likelihood (7b) with the derived proper non-informative prior (21), we get the posterior

$$p(\boldsymbol{\beta}|I) = \frac{1}{\pi}\left(\frac{3}{4}\right)^{k-2} \frac{\left|X^T X\right|_{\min}^{1/2}}{\|\mathbf{y}\|_{\max}^{k}}, \quad \boldsymbol{\beta} \in \text{Ellipsoid}. \tag{25}$$

For the regression coefficients $\boldsymbol{\beta}$, we get the following multivariate distribution

$$p(\mathbf{y}, \boldsymbol{\beta}|X, \sigma, M) = \left(\frac{3}{4}\right)^{k-2} \frac{\left|X^T X\right|_{\min}^{1/2}}{\pi \|\mathbf{y}\|_{\max}^{k}} \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{(\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta})}{2\sigma^2}\right]. \tag{26}$$

Integrating out the k unknown parameters $\boldsymbol{\beta}$ we are left with the following

marginal likelihood [4], that is, evidence value

$$
\begin{aligned}
p\left(\mathbf{y}|X,\sigma,M\right) &= \int p\left(\mathbf{y},\boldsymbol{\beta}|X,\sigma,M\right)d\boldsymbol{\beta} \\
&= \left(\frac{3}{4}\right)^{k-2}\frac{\left|X^TX\right|_{\min}^{1/2}}{\pi\left\|\mathbf{y}\right\|_{\max}^{k}}\frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\int\exp\left[-\frac{\left(\mathbf{y}-X\beta\right)^T\left(\mathbf{y}-X\boldsymbol{\beta}\right)}{2\sigma^2}\right]d\boldsymbol{\beta} \\
&= \left(\frac{3}{4}\right)^{k-2}\frac{\left|X^TX\right|_{\min}^{1/2}}{\pi\left\|\mathbf{y}\right\|_{\max}^{k}}\frac{\left(2\pi\sigma^2\right)^{k/2}}{\left|X^TX\right|^{1/2}}\frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\exp\left[-\frac{\left(\mathbf{y}-X\hat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y}-X\hat{\boldsymbol{\beta}}\right)}{2\sigma^2}\right]
\end{aligned}
\tag{27}
$$

where $\hat{\boldsymbol{\beta}} \equiv \left(X^TX\right)^{-1}X^T\mathbf{y}$ is the likelihood estimate of $\boldsymbol{\beta}$. Examining (6), we see that the evidence $p\left(\mathbf{y}|\mathbf{x},\sigma,M\right)$ may be deconstructed as

$$
p\left(\mathbf{y}|\mathbf{x},\sigma,M\right) = \frac{V_{\text{Post.}}}{V_{\text{Prior}}}L_{\text{Best}}, \tag{28}
$$

where $L_{\text{Best}}$is the best fit likelihood

$$
L_{Best} = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}}\exp\left[-\frac{\left(\mathbf{y}-X\hat{\beta}\right)^T\left(\mathbf{y}-X\hat{\beta}\right)}{2\sigma^2}\right]. \tag{29}
$$

$V_{\text{Post.}}$ is the volume of the posterior accessible region [1],

$$
V_{\text{Post.}} = \frac{\left(2\pi\sigma^2\right)^{k/2}}{\left|X^TX\right|^{1/2}} \tag{30}
$$

and $\mathrm{V}_{\text{Prior}}$ is the volume of the prior accessible region

$$
\mathrm{V}_{\text{Prior}} = \frac{\left\|\mathbf{y}\right\|_{\max}^{k}}{\left|X^TX\right|_{\min}^{1/2}}\left(\frac{4}{3}\right)^{k-2}\cdot\pi. \tag{31}
$$

Note that the posterior accessible region $V_{\text{Post.}}$ is an ellipsoid centered around the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ which lies inside the greater ellipsoid $\mathrm{V}_{\text{Prior}}$ of the prior accessible region centered at the origin. In Bayesian literature the ratio $V_{\text{Post.}}/\mathrm{V}_{\text{Prior}}$ is called the Occam factor. The Occam factor is equal to the factor by which M's hypothesis space collapses when the data arrive [1]. Looking at (28), we see that, in the specific case of equal prior probabilities for the models, that is, (4), Bayesian model comparison becomes a simple extension of maximum likelihood model selection. In the former the different models the best-fit likelihood values L_{Best} times their corresponding Occam factors $V_{\text{Post.}}/\mathrm{V}_{\text{Prior}}$ are compared, while in the latter only the best-fit likelihood values L_{Best} are compared.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Posterior predictions on river discharges

D.J. DE WAAL* – University of the Free State, Bloemfontein, South Afrika

Abstract. The late Jan van Noortwijk (JvN) made valuable contributions in many areas such as Reliability, Risk management, Maintenance modelling, Applications to Decision theory and more. His contributions to model river discharges for flood prevention (van Noortwijk *et al.*, [1, 2] and others) are of interest to forecast river stream flow. The posterior predictive densities for several distributions, which can be considered as candidates to model river discharges, were derived using Jeffreys prior. The Jeffreys prior was derived for these distributions by careful algebraic derivations of the Fisher information matrix. The posterior predictive density is the way we believe to follow for predicting future values once the best model is selected. Van Noortwijk *et al.* [1, 2] proposed Bayes weights for selecting the best model. The advantage of the posterior predictions over substituting the estimates of the parameters in the quantile function is discussed for a special case. A further application under regression in the lognormal model with the Southern Oscillation Index (SOI) as independent variable, is shown for the annual discharge of the Orange River in South Africa. It implies the prediction of the SOI at least one year ahead through an autoregressive time series.

1 INTRODUCTION

Van Noortwijk *et al.* [1], [2] considered several distributions as possible candidates to model river discharges to predict future floods. A Bayesian approach was followed using the Jeffreys prior in each case. With careful algebraic manipulations, he derived the Fisher information matrix for these distributions, namely Exponential, Rayleigh, Normal, Lognormal, Weibull, Gamma, Generalised Gamma, Inverted Gamma, Student t, Gumbel, Generalised Gompertz, Generalised Extreme Value, Pareto and Poisson distributions. Future river discharges were predicted through the posterior predictive distribution, which were derived for the above cases and model selection were discussed through the Bayes Factor. All these techniques can be considered sound. The Bayes paradigm allows for other relevant sources of information instead of only measured values, but if this information puts too much weight on the posterior distribution, a non-informative prior such

*corresponding author: Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein 9301, South Africa; telephone: +27-51 4012311, fax: +27-51 4442024 e-mail: deWaalDJ.SCI@ufs.ac.za

as the Jeffreys which JvN used, can be a good choice. He followed the Bayesian route to predict future values by deriving the posterior predicted distribution for each of the above distributions. In Section 2, the difference between the Bayesian predictive approach and the 'plug-in' method where the estimates of the parameters are just plug into the quantile function, is discussed for the Exponential case. In Section 3 the lognormal is applied for predicting river discharges. This is a case where the predictive posterior distribution is log-t and numerical integration is necessary to obtain predictive quantiles. In Section 4 the prediction of the Southern Oscillation Index (SOI), which was introduced as a variable to improve on the predictions, is discussed. Section 5 is devoted to a discussion on validating the predictions.

2 POSTERIOR PREDICTION

The advantages of the posterior predictive approach above the 'plug-in' method were the quantile function is estimated, are considered through simulating data from an exponential distribution. One big advantage is that extremes may be scarce in the data, but can be predicted using the Bayesian approach. The box plots shown in Figure 1 were drawn from 500 samples $x_1, ..., x_n$ of sizes $n = 5$ and $n = 10$ from an exponential distribution with location parameter $\lambda = 10$ and distribution function (df) given by

$$F(x) = 1 - e^{-x/\lambda}, \quad x > 0. \tag{1}$$

The box plots in Figure 1 (a and c) show the distribution of the estimated quantiles from the quantile function

$$Q(p) = -\lambda \log(1-p), 0 < p < 1. \tag{2}$$

p is chosen as $i/(n+1)$, $i = 1, ..., n$ and λ is estimated by $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$. By plugging in the estimate of λ in 2, the estimate of the quantile function is obtained and this is referred to as the "plug-in" method. The box plots in Figure 1 (b and d) are the posterior predictive distributions simulated from the 500 samples. The posterior predictive quantile function is given by

$$Q^{PRED}(p) = n\bar{x}\{(1-p)^{-1/n} - 1\}. \tag{3}$$

It follows that as $n \to \infty$, 3 approaches

$$Q^{EST}(p) = \bar{x} \log(1-p). \tag{4}$$

Equation 4 follows from the posterior predictive distribution function (van Noortwijk *et al.* , [2], page $A-4$)

$$P(X > x_0|x) = \left(1 + \frac{x_0}{n\bar{x}}\right)^{-n}. \tag{5}$$

Equation 5 is the posterior predictive survival function exceeding a future x_0 and is recognised as that of a Generalised Pareto. The Jefrreys prior $\pi(\lambda) \propto 1/\lambda$, is used as the prior on λ. The posterior of λ becomes an Inverse Gamma$(n, n\bar{x})$.

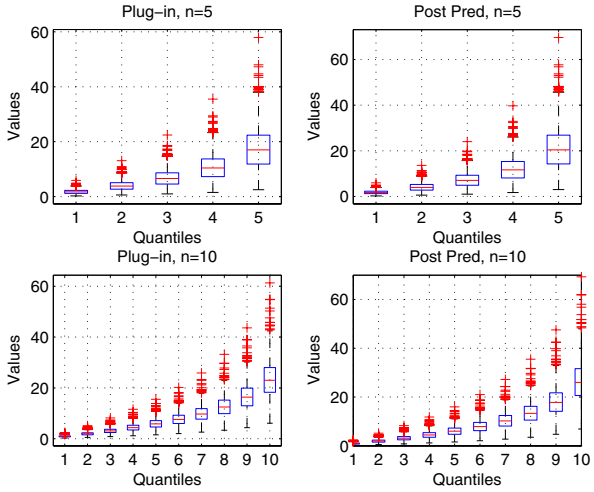


FIGURE 1. Boxplots comparing distributions of estimated quantiles (a, c) with predicted quantiles (b, d) for different quantiles and sample sizes

We notice from the figures that the larger predictive quantiles show much heavier tails and larger predicted values than the estimated quantiles. This can cause severe under estimation especially for small sample sizes. As the sample size increases, the predictive quantiles approach the estimated quantiles. It is therefore advisable to take the predictive road always. In many cases explicit expressions are not possible such as the above, but one can always do simulations.

3 THE LOGNORMAL MODEL

We will now consider the prediction of the annual volume inflow into the Gariep Dam from the discharges of the Orange River in South Africa. This is important for ESKOM (main supplier of electricity in South Africa) to be able to manage the generation of hydro power at the dam wall without spilling water over the wall and to maximize their power generation through the four turbines.

3.1 Orange River stream flow data

Figure 2 shows the annual volume discharges $x = (x_i, i = 1, ..., n = 37)$ of the Orange river during $1971 - 2007$ in million m^3.

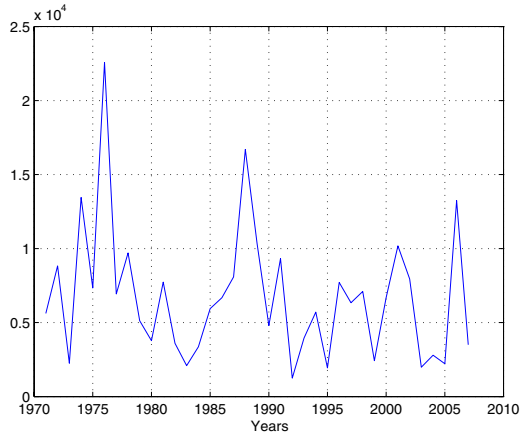


FIGURE 2. Annual volume inflow in million cubic meter to gariep Dam during $1971 - 2007$

The mean annual inflow is $6.7354 \times 10^3\ m^3$ and standard deviation $4.4652 \times 10^3\ m^3$. Assuming the annual inflows as independent, a lognormal, $LN(\mu, \sigma)$, distribution is fitted due to the heavy tail indicated by the data. The independence is assumed, since the autocorrelation between successive years is quite small 0.04, and reaches a maximum of 0.4 if the lag is 11 years. The 11 year cycle corresponds to the sunspot cycle of 11 years which is well known among Astronomers. (See Matlab demo [6] on Fast Fourier Transforms). Fitting a LN to the data with $\hat{\mu} = 8.6094$ and $\hat{\sigma} = 0.6698$, we obtain a fairly good fit according to QQ-plots comparing predicted stream flows with the observed shown in Figure 3. Van Noortwijk *et al.* [2] discussed the use of Bayes factors to compare the fit of several models and to select the best model. This will briefly be discussed in Section 3.2.

3.2 Model selection and Goodness of fit

Model selection

Comparing two models, the posterior odds can be written as the product of the Bayes factor times the prior odds. To select the best model among k specified models with equal prior weights, van Noortwijk *et al.* [2] considered the calculation of the posterior probability (refered to as the Bayes weights). The model with the largest posterior probability is chosen as the best. The derivation of the marginal density $\pi(x|M_i)$ cannot be obtained explicitly in many cases and numerical integration has to done. Van Noortwijk *et al.* [1, 2] put in effort to show that posterior weights can be used to decide on the best model. They remarked that more than one model seems to fit using goodness-of-fit tests and proposed the calculation of Bayes weights.

In general, the lognormal is considered a good candidate for modeling river discharges and therefore we choose the lognormal for modeling the stream flow of the Orange river without considering other models.

Goodness of fit of the lognormal

To test the goodness of fit of the lognormal to data in Figure 2, we use the predictive approach to predict a set of n observations and compare the predicted with the observed data in a QQ-plot. Gelman *et al.*, [3], page 69 and van Noortwijk *et al.*, [2]) showed that under the non informative Jeffreys prior on μ and σ, namely

$$\pi(\mu, \sigma) \propto 1/\sigma, \tag{6}$$

the posterior predictive distribution of $Y = \log(X_0)$ becomes a t distribution with $n - 1$ degrees of freedom and parameters

$$\bar{y} = \sum_{i=1}^{n} \log x_i / n. \tag{7}$$

and

$$S_y^2 = \frac{(n+1)^2}{n^2} \sum_{i=1}^{n} (\log x_i - \bar{y})^2. \tag{8}$$

The df of a future X_0 becomes

$$P(X_0 < x_0 | x) = t_{n-1}(\log(x_0) - \bar{y})/S_y \tag{9}$$

$n = 37$ observations were predicted using the $t(n-1, \bar{y}, S_y)$ after taking the exponentials of the t-predictions and compare with the sorted original observations in a QQ-plot. This was repeated 500 times and the smallest correlation between the observed and predicted values were 0.8417. Repeating this by plugging in the estimates in the lognormal and calculating the correlation between observed and estimated quantiles, the smallest correlation of 0.7828 is obtained. The means of the correlations from these two types of QQ-plots were both high although for the plug-in case there were smaller correlations. We can therefore be quite satisfied that a lognormal is a good choice. From the 500 predictions of the 37 years inflow, the mean prediction is 6781.8×10^6 m^3 with 95% highest posterior density (hpd) region $(1268 \times 10^6, 23624 \times 10^6)$. The maximum annual inflow observed during the 37 years, is 22571×10^6 and the minimum is 1244×10^6. The observed trend in the data appears to be quite possible from the predictions and is not significant. To improve on these predictions, we looked for other variables that can be introduced into the model and discovered the Southern Oscillation Index as a possible indicator. We will explore this further in the next section.

3.3 Introducing the SOI as an independent variable

The Southern Oscillation Index (SOI) which measures the difference in pressure between Darwin and Tahiti is a well known indicator of rainfall in the southern hemisphere. A number of studies were done on this phenomenon and it is found to be an indicator of the rainfall in the Southern hemisphere. We compare the annual volume of stream flow into the Gariep dam with the SOI of October the corresponding to previous year. The annual stream flow was correlated with different months and lags and October of the corresponding to previous year was selected as the month, which has the highest correlation with the year inflows. A linear regression for predicting the inflow (Y) given the SOI of October (X) on the 37 years $1970 - 2006$ is considered, namely $E(Y) = a + bX$. The estimates of a and b are 6952.7 and 215.6 respectively. Substituting the X values (SOI) in the equation, estimates of the stream flow are obtained which are shown in Figure 3 together with the true inflows. The correlation between the estimated and true inflows is $r = 0.5062$. This is quite remarkable.

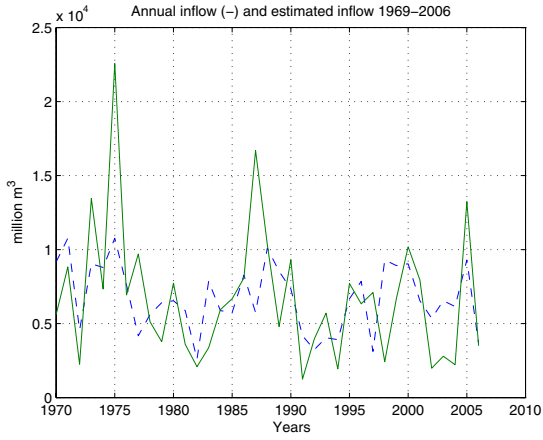


FIGURE 3. Comparing true inflows (-) with estimated inflows (- -) using October SOI of previous year

With this method, we can predict the annual volume of inflow for the following year given the SOI for October the previous year. This implies we have to predict the SOI for the next year. This will be addressed in Section 4. The trend of the inflow volume over the 37 years $1970 - 2006$ is showing a decrease (Figure 2) of 3287 million m^3, it is 88 million m^3 per year. The mean annual inflow is 6735.4 million m^3. To make a long term prediction like 5 or 10 years ahead, ignoring the SOI, we can say that over the next 5 years, the inflow on average will be 445 million m^3 less and over 10 years 890 million m^3 less. We indicated above that the SOI for October

has an impact on the annual inflow Y. We will explore this relationship further under the lognormal model with a regression on the SOI to see if we can improve on the above model. Let $y = X\beta + u$ where $y = \log(Y(n,1))$, $X = [\mathbf{1} \cdot \text{SOI}(oct)](n,2)$, $\beta = [\beta_1\beta_2](2,1)$, where $\mathbf{1}$ is a $(n,1)$ vector of ones and the elements of $u(n,1)$ are distributed independently $N(0,\sigma^2)$. It follows that the predictive density of a future $\tilde{y}(q,1)$ given the corresponding covariate matrix $\tilde{X}(q,2)$, is a multivariate t_v distribution (Zellner, [4], pp $72-74$) with $v = n-2$ degrees of freedom, mean $\tilde{X}\beta$ and covariance matrix $\frac{v}{v-2}s^2(I - \tilde{X}M^{-1}\tilde{X}')^{-1}$, $s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/v$ and $M = X'X + \tilde{X}'\tilde{X}$. The model above can also be considered as hierarchical within a latent process with $X\beta + u$ a latent variable. (See Steinback *et al.*, [5]). An advantage of this approach is that the influence of the latent process on model assumptions can be checked separately. Further posterior predictive p-values can be used to test assumptions on different levels. We will however not proceed with this further now. If $q = 1$, we predict one year ahead, then

$$\frac{\tilde{y} - \tilde{X}\hat{\beta}}{s/\sqrt{(1 - \tilde{X}M^{-1}\tilde{X}')}} \sim t_v.$$

Looking at one year ahead, we can simulate t-values. Then $T = \exp(t)$ will be the predicted values. If we repeat this simulation a number of times for a given October SOI, we can calculate the median(T) with lower and upper quartiles. Repeating this simulation for varying SOI values, we are able to construct a table, showing the predicted median inflow with lower and upper quartiles for a given October SOI. From the inflow data for $1970 - 2006$ and the SOI value for October of the previous year, we can now predict the inflow for a year ahead. We calculated $M = (38 - 17.3; -17.34395)$, $s = 0.5768$ and $\hat{\beta} = (8.6434, 0.0337)$. The number of degrees of freedom is $v = 37 - 2 = 35$. Suppose a positive index of 6 for October, then $\tilde{X} = [1; 6]$ and the predictive t-values can be simulated. From say 1000 simulated t-values, the median (or mean) can be calculated as the predictive value together with quartiles or any hpd region. We found that for a SOI of 6, we predicted an annual inflow of 6991 million m^3. The prediction without the SOI factor, was given in Section 3.2 as 6781 million m^3. Notice that the inter quartile range in the regression model is also much smaller. Figure 4 shows the predictions with the inter quartile range. Figure 5 shows a set of simulated predictions given a SOI $= 6$ with a histogram of the predicted values and Figure 6 shows box plots of the simulated predictions at different SOI values ranging from -20 to 20. To predict say 5 years ahead, we need to use the multivariate t model with predicted SOI values. We predicted the observed 37 years inflows that we observed given the corresponding SOI values and the predictions were almost similar to the predictions we made one year ahead. The gain is so small that it is not worth to follow the multivariate approach. There is a significant negative auto correlation of -0.2255 found between a 2 year lag on the December SOI's and a positive

correlation of 0.5450 found between the December SOI's and the October SOI's. To predict the October SOI values are therefore not a simple matter and we have to dig into the literature on this issue further.

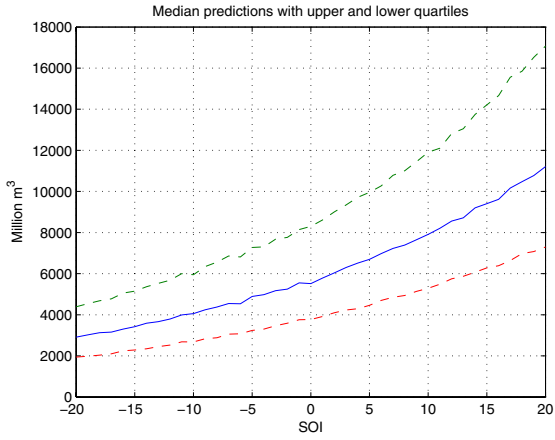


FIGURE 4. Predicted annual median inflows (-) with upper and lower quartiles (--)

4 PREDICTING THE SOI

We observed in the previous section that the annual inflow prediction for the next year can be improved by introducing the SOI for October of that year. This means that we need to predict the SOI which is not that simple. There exists a vast literature on the SOI and models to predict the SOI. Dr T Landscheidt, a expert on SOI, commented that the SOI cannot be predicted more than 12 months ahead. Dr Wasyl Drosdowsky at the Bureau of Meteorology Research Centre (BMRC) in Australia developed time series methods to project the SOI into the future. Dr Neville Nicholls at BMRC is also an expert on SOI and his article in the web site `http://www.abc.net.au/science/slab/elnino/story.htm` is worth reading. The National Climatic Data Center (NOAA) in the USA is also a valuable source of information on future SOI predictions. Their web site is `http://www.ncdc.noaa.gov/oa/climate/research/2008/enso-monitoring.html`. According to their forecast, we can expect a mild positive index for October 2008. From Figure 5, it means we can expect an annual inflow of 6500 million m^3 with quartiles $(4300, 10000)$. Figure 7 shows the October SOI index for the period $1876 - 2007$.

The SOI is affected by various climatic factors of which winds play an important role. The sun spots which have a cycle of 11 years (see The

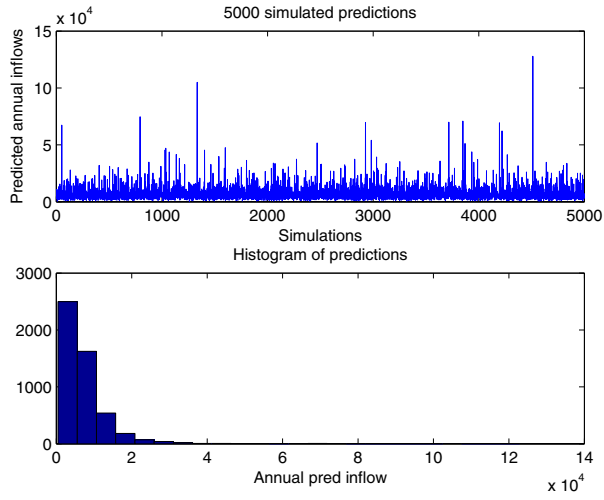


FIGURE 5. A simulation of 5000 predicted inflows with SOI = 6 and a histogram of simulated predictions with SOI = 6

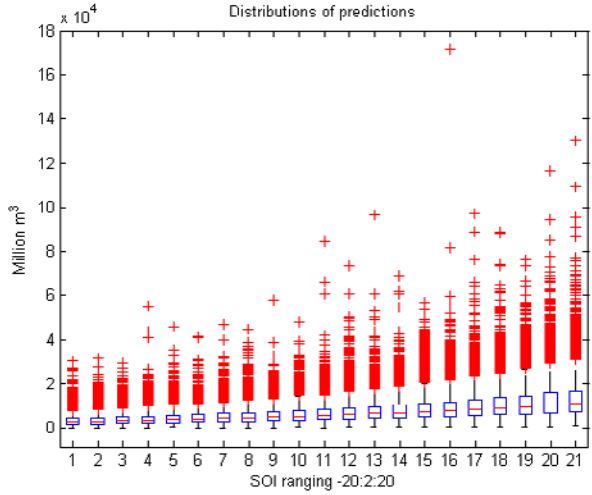


FIGURE 6. Boxplots of 5000 simulated predictions from the log-t distribution for different October SOI values ranging from −20 to 20 with steps of 2

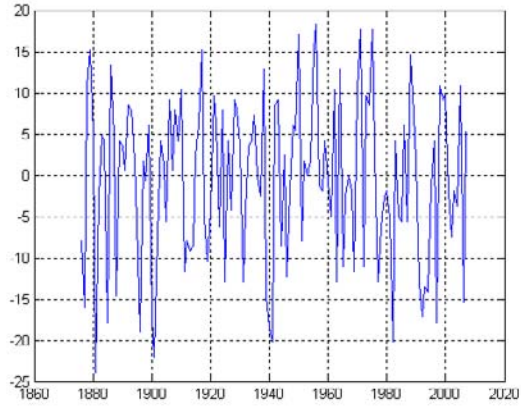


FIGURE 7. SOI of October 1876 – 2007

MathWorks, [6], are also important indicators. Predicting the SOI monthly, we investigated the monthly data using spectral techniques, but it turned out to be not worth the effort to predict it due to too much noise. Our results coincide with that of Mills [7] who discussed the prediction of the North Atalantic Oscillation (NAO). He also came to the conclusion that the monthly time series models investigated, explains less than 15% of the variation in the NAO index. Salisbury and Chepachet [8] used the Empirical Mode Decomposition (EMD) method and claims an improvement on SOI predictions. Since we related the annual inflow to the October SOI, we are only interested in predicting the October SOI. Applying a spectral analysis on the last 56 years 1952 – 2007, and using 7 harmonics, we are able to declare 64% of the variation. The reason why 56 years are taken, is that the earlier SOI do not seem to be very reliable and therefore we consistently use only 56 years of data. Figure 8 shows a strange phenomenon. The correlation between the first 56 observed October SOI values and those estimated from the Fourier series, indicated on the graph against 1931, is below 0.5 and it stays fairly the same for the next consecutive 56 years till 1955 and then we have a sudden drop to 0.33. From 1961 it gradually increases till 0.8 in 2008. We therefore decided to stick to 56 years for the forecasts. We will investigate the prediction of the SOI further.

5 VALIDATING PREDICTIONS

5.1 Validation of log-t predictions

To validate the predictions from the log-t model, we compare the true inflows with the predicted for the years 1970 – 2006 given the October SOI values.

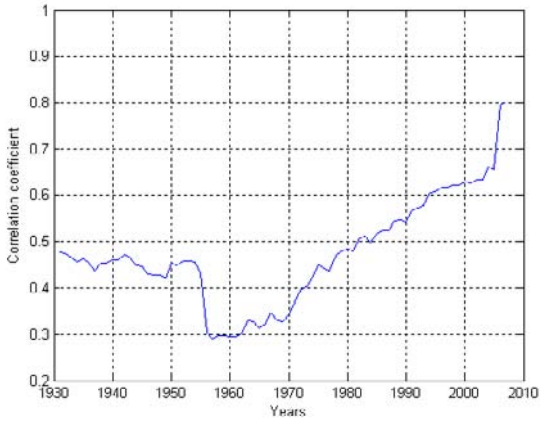


FIGURE 8. Correlations between SOI predictions for $1931 - 2007$ and the observed based on the 56 previous years data

We managed to increase the correlation from 0.5062 (see Section 2.3) to 0.5441 with this model. A box plot of the predicted simulated distributions is shown in Figure 9.

5.2 Validating the SOI predictions

Applying a spectral analysis on the last 56 years $1952 - 2007$, and using 7 harmonics, we are able to declare 64% of the variation. Figure 10 shows the October SOI data with the estimated October SOI. A comparison of the estimates with the data through a box plot, is shown in Figure 11. The first column shows a distribution of the data, the second column the distribution of the estimates. The predicted values for 2008 to 2012 are 4, -6, -3, 7 and 8 respectively.

5.3 Validating the inflow predictions one year ahead

We can now proceed to predict the inflows after predicting the SOI for October of the previous year with credibility intervals. Comparing the predicted one year ahead with the true annual inflows and calculating the success rate by expressing the number of times the true inflows falls between the lower and upper values, we got 35%. This success rate is based on the accuracy of the SOI predictions for October and will be further investigated. We are still doing better with this method than applying the regression model discussed in Section 2.3.

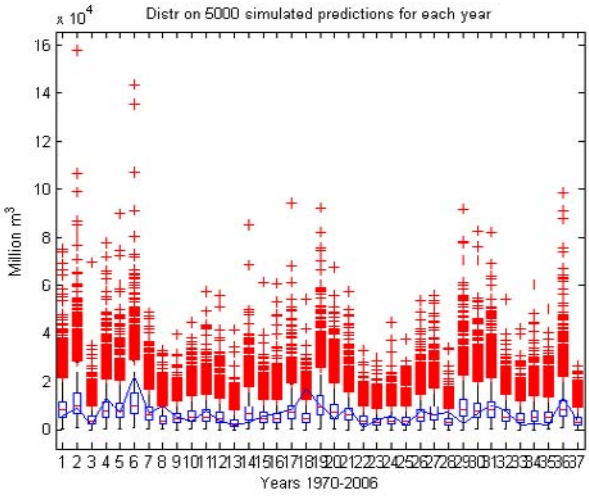


FIGURE 9. Boxplots of the simulated predictions from the log-t for the years $1970 - 2006$ and the true annual inflows (-)

6 CONCLUSION

In conclusion: Jan van Noortwijk was on the way to more useful contributions in various fields and I know that it was his intention to publish the work referred to in this presentation in book form. It is a pity that he could not reach that goal. His section on model selection contains some new ideas, some of which can be debated. One question is: How important is the number of distributions that are considered as possible candidates for modeling river discharges in the Bayesian weights? In this presentation only one model, namely the lognormal, is selected and tested for acceptance as an appropriate model. Once this obstacle is out of the way, the prediction of future observations becomes the issue. It has been shown how important it is to introduce additional information such as the SOI and how it is introduced. From the predicted SOI values 1 to 5 years ahead, the inflows can be predicted. For example, suppose the SOI for 2008 is 4 and therefore from Figure 4, we get a predicted annual inflow for 2009 of approximately 6300 million m^3. The confidence bounds are $(4251, 9608)$. Matlab programs were developed for making these predictions.
To summarize the findings:

1. An alternative to a regression model through fitting a time series model is suggested by considering the annual volume of inflow. This is modeled through a lognormal distribution under certain assumptions. From a Bayesian perspective, the predictive density is derived and predictions on future annual inflows are made.

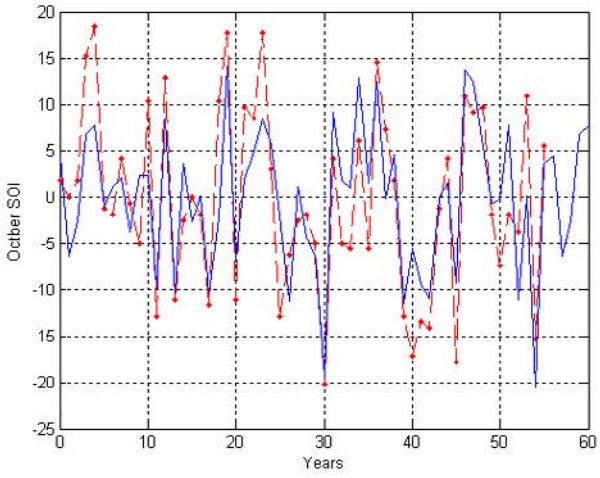


FIGURE 10. October SOI data 1952 – 2007 (.–) and estimates (*-) with predictions for 2008 – 2012

2. These predictions are improved by introducing the SOI for October of the previous year as an independent variable in the regression model.

3. The prediction of the SOI needs to be investigated further. At this stage a Fourier series is fitted.

4. The joint distribution of Inflow and October SOI has been considered and gives further insight into the behavior of the variables. This will be prepared for a future communication.

Acknowledgments

I want to express appreciation for assistance of colleagues dr. Pieter van Gelder and Yuliya Avdeeva of the Delft University of Technology for a contribution they made, dr. Martin van Zyl, mr. Leon Fourie, mr. S. van der Merwe , Mrs. Christelle Mienie and Bernice Pfeifer of the department Mathematical Statistics and Actuarial Science at the University of the Free State. To ESKOM my appreciation for supplying data and for financial support.

Bibliography

[1] J. M. Van Noortwijk, H. J. Kalk, M. T. Duits, and E. H. Chbab. The use of bayes factors for model selection in structural reliability. In Structural Safety and Reliability, editors, *In R. B. Corotis, G. I. Schuëller and M. Shinozuka*, volume Proceedings of the 8th International Conference on Structural Safety and Reliability, ICOSSAR, Lisse, Balkema, 2001.

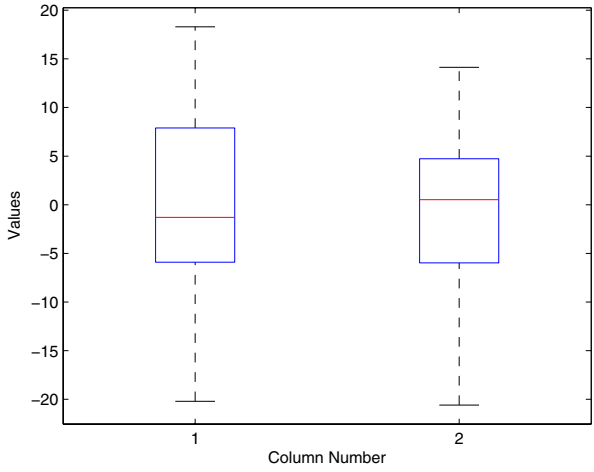


FIGURE 11. Boxplot of October SOI for 1952 − 2007 (first column) and the estimated using 7 harmonics

[2] J. M. van Noortwijk, H. J. Kalk, M. T. Duits, and E. H. Chbab. Bayesian statistics for flood prevention. Technical Report PR280, Ministry of Transport, Public Works and Water Management, Institute forInland Water Management and Waste Water Treatment (RIZA), and HKV Consultants, Lelystad, The Netherlands, 2003.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[4] A Zellner. *An Introduction to Bayesian Inference in Econometrics*. Wiley, 1971.

[5] G. H. Steinbakk and G. O. Storvik. Posterior predictive p-values in bayesian hierarchical models. *Scandinavian Journal of Statistics*, 36:320–336, 2009.

[6] Matlab statistical toolbox, demo program fft, 2007. URL `www.mathworks.com`.

[7] T. C Mills. Is the north atlantic oscillation a random walk? a comment with further results. *International Journal of Climatology*, 24:377–383, 2004.

[8] J. I. Salisbury and M. Wimbush. Using modern time series analysis techniques to predict enso events from the soi time series. *Nonlinear Processes in Geophysics*, 9:341–345, 2002.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

The lessons of New Orleans

J.K. Vrijling* – Delft University of Technology, Delft, the Netherlands

Abstract. End of August 2005 the flood defences of New Orleans were hit by hurricane Katrina. It quickly became apparent that they could not withstand this force of nature. The three bowls of the city were flooded. Over a thousand people lost their lives and the total damage exceeded $20 billion US. What can we learn from this disaster? Can the process of understanding be supported by mathematics? Is it possible to draw conclusions with the help of mathematics that can help to avoid a repeat of this tragedy?

Two years after the disaster no decision has been taken about the required level of protection. This is a mathematical decision problem where the increasing cost of protection is equated with the reduced risk (probability × consequence) of flooding. Where the sum of the cost of protection and the present value of the risk reaches a minimum, the optimal level of protection is found. Along this line of reasoning the level of flood protection of the Netherlands was decided in 1960. However today some think that an insurance against the consequences of flooding is to be preferred over spending money on a flood defence system that will never be absolutely safe. Others judge it necessary to prepare the evacuation in case of a flood because perfect safety by flood protection is unattainable. Mathematics shows that both options are probably no alternative to optimal prevention.

1 INTRODUCTION

End of August 2005 the flood defenses of New Orleans were hit by hurricane Katrina. It quickly became apparent that they could not withstand this force of nature. The three bowls of the city were flooded. Over a thousand people lost their lives and the total damage exceeded $20 billion US.

What can we learn from this disaster? Can the probabilistic design models that were developed in the last decades help to improve the insight? The answer seems affirmative.

The simple lesson that a flood defense system is a series-system became very clear. The weakest link decides the overall safety of the system. At some places the defense was non-existent, so flooding was in fact a certainty with an above average hurricane. Additionally, some parts were three feet

*corresponding author: Hydraulic Engineering and Probabilistic Design Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5, 2600 Delft, The Netherlands; telephone: +31-(0)15 27 85278, e-mail: j.k.vrijling@tudelft.nl

short due to confusion about the datum. Finally, parts of the system were pushed backwards and failed before the storm surge level reached the crest of the wall.

Two years after the disaster no decision has been taken about the required level of protection. This is a decision problem where the increasing cost of protection is equated with the reduced risk (probability x consequence) of flooding. Where the sum of the cost of protection and the present value of the risk reaches a minimum, the optimal level of protection is found. The level of flood protection of the Netherlands was decided in 1960 on this basis.

However today some think that an insurance against the consequences of flooding is to be preferred over spending money on a flood defense system that will never be absolutely safe. Others judge it necessary to prepare the evacuation in case of a flood because perfect safety by flood protection is unattainable. Probability theory shows that both options are generally no alternative to optimal prevention.

2 THE FLOOD DEFENSE SYSTEM AS A SERIES-SYSTEM

The last decades probabilistic design methods instilled the awareness, that the probability of exceedance of the design water level, the design frequency or the reciprocal of the return period is not an accurate predictor of the probability of flooding. Traditionally the dike crest exceeds the design water level by some measure, thus the probability of overtopping is smaller than the design frequency. But water logging may lead to slide planes through the dike or piping may undermine the body of the dike, with sudden failure as a consequence. Both are not accounted for in the design frequency. In short there are more failure mechanisms that can lead to flooding of the polder than overtopping (see Figure 1). Human failure could prohibit the timely closure of sluices and gates before the high water. Moreover the length of the dike ring has a considerable influence. A chain is as strong as the weakest link. A single weak spot may determine the actual safety of the dike ring (Figure 2).

The probabilistic approach aims to determine the probability of flooding of a polder and to judge its acceptability in view of the consequences. As a start the entire water defense system of the polder is studied. Typically this system contains sea dikes, dunes, river levees, sluices, pumping stations, high hills, etc. (Figure 2).

In principle the failure and breach of any of these elements leads to flooding of the polder. The probability of flooding results thus from the probabilities of failure of all these elements. Within a longer element e.g. a dike of 2 km length, several independent sections can be discerned. Each section may fail due to various failure mechanisms like overtopping, sliding, piping, erosion of the protected outer slope, ship collision, bursting pipeline, etc. The relation between the failure mechanisms in a section and the

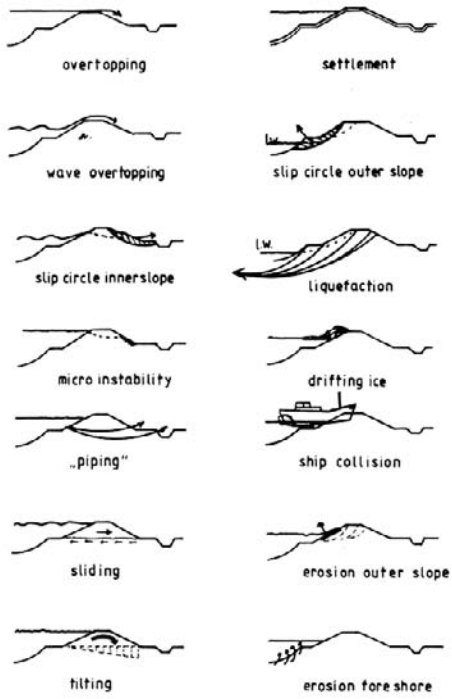


FIGURE 1. Failure modes of a dike; from [1]

unwanted consequence of inundation can be depicted with a fault tree as shown in Figures 2 and 3 in which the following notation is used: R_i the resistance of section I, e.g. h the height of the dike, B the width of the dike or D the size of the revetment block and S_i the solicitation, e.g. wl the water levels and Hs the wave heights in front of the dike.

The failure probabilities of the mechanisms are calculated using the methods of the modern reliability theory like Level III Monte Carlo, Level II advanced first order second moment calculations (see [2, 3, 4, 5, 6] or Van Gelder [7] for a complete overview).

The experience in New Orleans proved that other mechanisms than overtopping contributes to the failure probability (ASCE [8]). Along the 17-th Street Canal any sign of overtopping was lacking, but the flood pushed the wall backwards due to failure of the sub-soil. At the London Avenue Canal massive piping led to failure of the concrete floodwall without overtopping. The protective sheet pile wall at the Ninth Ward was overtopped and lost stability completely. The possibility to treat the human failure to close for instance a sluice in conjunction with structural failure is seen as a considerable advantage of the probabilistic approach (Fig. 4). Nowak and

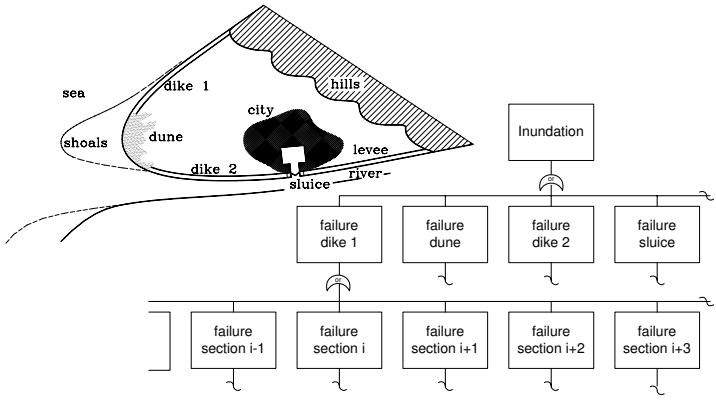


FIGURE 2. Flood defense system and its elements presented in a fault tree

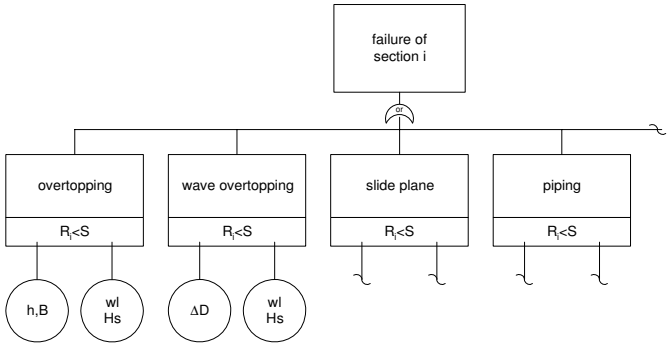


FIGURE 3. A dike section as a series system of failure modes

Collins [3] devote attention to this issue. In New Orleans only one of hundred flood gates was left open. However other human errors in establishing chart datum and in design contributed to the disaster.

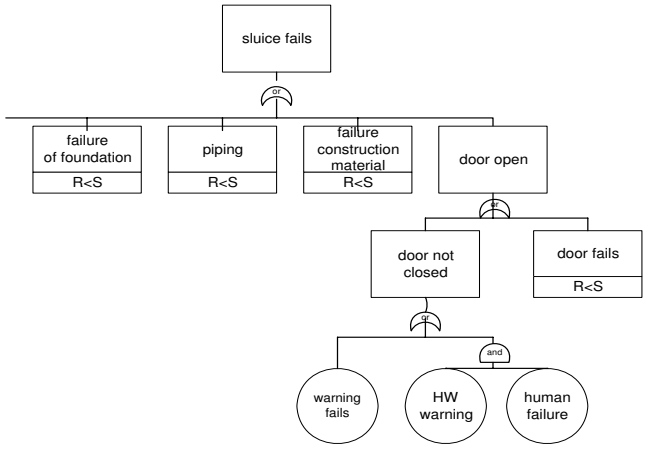


FIGURE 4. The sluice as a series system of failure modes

Correlations between failure modes and correlations between different dike sections have to be taken into account. Techniques are described in for instance Hohenbichler and Rackwitz [2]. In the reliability calculations all uncertainties should be dealt with. Three classes are discerned. The intrinsic uncertainty is characteristic for natural phenomena. Model uncertainty describes the imperfection of the engineering models in predicting the behaviour of river flows, dikes and structures. The comparison of predictions and observations provides an estimate of this uncertainty. Statistical uncertainty is caused by the lack of data. These data are used to estimate the parameters of the probability distributions depicting the intrinsic uncertainty. Because all uncertainties are included in the calculations of the failure probability the latter is not singly a property of the physical reality but also of the human knowledge of the system (Blockley [9], and Stewart and Melchers [4]).

The result is that the safety of the dike system as expressed by the calculated probability of flooding can be improved by strengthening the weakest dike but also by increasing our knowledge. The result of the calculated probability of flooding of the polder is presented in Table1.

The last column of the table shows immediately which element or section has the largest contribution to the probability of flooding of the polder

Section	Overtopping	Piping	Etc.	Total
dike section 1.1	$p_{1.1}$(overtop.)	$p_{1.1}$(piping)	$p_{1.1}$(etc.)	$p_{1.1}$(all)
dike section 1.2	$p_{1.2}$(overtop.)	$p_{1.2}$(piping)	$p_{1.2}$(etc.)	$p_{1.2}$(all)
etc.	...	...	...	...
dune	p_{dune}(overtop.)	p_{dune}(piping)	p_{dune}(etc.)	p_{dune}(all)
sluice	p_{sluice}(overtop.)	p_{sluice}(piping)	p_{sluice}(etc.)	p_{sluice}(all)
total	p_{all}(overtop.)	p_{all}(piping)	p_{all}(etc.)	p_{all}(all)

TABLE 1. Table with the contributions to the overall probability of inundation

under study. Inspection of the related row reveals which mechanism will most likely be the cause. Thus a sequence of measures can be defined which at first will quickly improve the probability of flooding but later runs into diminishing returns.

3 THE ACCEPTABLE PROBABILITY OF FAILURE

One of the tasks of human civilizations is to protect individual members and groups to a certain extent against natural and man-made hazards. The extent of the protection was in historic cases mostly decided after the occurrence of the hazard had shown the consequences. The modern probabilistic approach aims to give protection when the risks are felt to be high. Risk is defined as the probability of a disaster i.e. a flood related to the consequences. As long as the modern approach is not firmly embedded in society, the idea of acceptable risk may, just as in the old days, be quite suddenly influenced by a single spectacular accident like the inundation of New Orleans or an incident like the non-calamitous threats of the Dutch river floods of 1993 and 1995.

The estimation of the consequences of a flood constitutes a central element in the modern approach. Most probably society will look to the total damage caused by the occurrence of a flood (Vrijling [10]). This comprises a number of casualties, material and economic damage as well as the loss of or harm to immaterial values like works of art and amenity. Also the loss of trust in the water defence system is a serious, but difficult to gauge effect. However for practical reasons the notion of risk in a societal context is often reduced to the total number of casualties using a definition as: "the relation between frequency and the number of people suffering from a specified level of harm in a given population from the realization of specified hazards". If the specified level of harm is limited to loss of life, the societal risk may be modelled by the frequency of exceedance curve of the number of deaths, also called the FN-curve [10].

The consequence part of a risk can also be limited to the material damage expressed in monetary terms as the Dutch Delta Committee did in 1960. It should be noted however, that the reduction of the consequences

of an accident to the number of casualties or the economic damage may not adequately model the public's perception of the potential loss. The schematisation clarifies the reasoning at the cost of accuracy.

The problem of the acceptable level of risk can be elegantly formulated as an economic decision problem. The expenditure I for a safer system is equated with the gain made by the decreasing present value of the risk. The optimal level of safety indicated by P_f corresponds to the point of minimal cost.

$$\min(Q) = \min(I(P_f) + \mathrm{PV}(P_f \cdot D)),$$

where:

Q = total cost
PV = present value operator
D = total damage given failure

If, despite ethical objections, the value of a human life is rated at d according to [11], the amount of damage is increased to:

$$P_{d|f_i} \cdot N_i \cdot d + D,$$

where:

N_i = number of inhabitants in polder i
$P_{d|fi}$ = probability of drowning given failure

This extension makes the damage an increasing function of the expected number of deaths. The valuation of human life is chosen as the present value of the nett national product per inhabitant. The advantage of taking the possible loss of lives into account in economic terms is that the safety measures are affordable in the context of the national income (see also Vrijling and Van Gelder [11]).

Omitting the value of human life, the decision problem as formulated by the Delta Committee [12, 13] is given below. The investment $I(h)$ in the protective dike system is given as a function of the crest level h by:

$$I(h) = I_0 + I_1(h - h_0),$$

where:

I_0 = initial cost
I_1 = marginal cost
h_0 = existing dike level

The annual probability of exceedance of the crest level of the dike is given by an exponential distribution:

$$1 - F(h) = e^{-\frac{h-A}{B}}.$$

The risk of inundation is equal to the probability of exceedance of the dike crest times the damage D in case of inundation.

$$Risk = e^{-\frac{(h-A)}{B}} \cdot D.$$

Because the risk occurs every year the present value of the risk over an infinite period has to be taken into account:

$$\mathrm{PV(Risk)} = \sum_{i=1}^{\infty} e^{-\frac{(h-A)}{B}} \frac{D}{(1+r)^i} = e^{-\frac{(h-A)}{B}} \frac{D}{r},$$

where r is the discount rate. The total cost is the sum of the investment and the present value of the remaining risk that is accepted;

$$\mathrm{TC}(h) = I_0 + I_1(h - h_0) + e^{-\frac{(h-A)}{B}} \frac{D}{r}.$$

Differentiating the total cost with respect to the decision variable h and equating the derivative to 0 gives an elegant result

$$\frac{\partial \mathrm{TC}(h)}{\partial h} = I_1 - \frac{1}{B} e^{-\frac{(h-A)}{B}} \frac{D}{r} = 0$$

$$p_f^* = e^{-\frac{(h-A)}{B}} = \frac{I_1 B r}{D}$$

The last expression shows that the acceptable probability increases with the marginal cost of dike construction, with the standard deviation of the storm surge level B and the rate of interest. It decreases with the damage that will occur in case of an inundation.

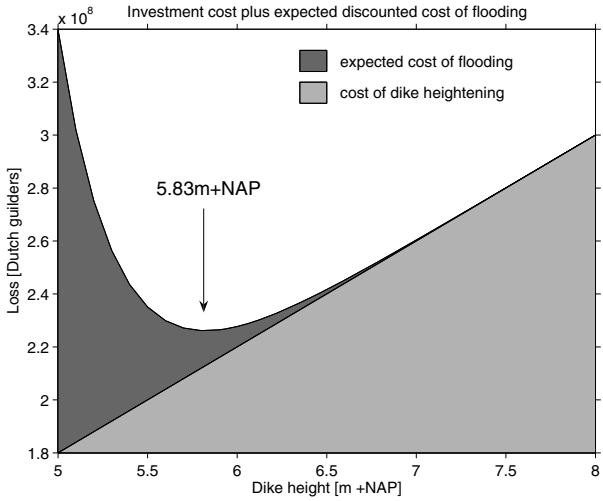


FIGURE 5. The economically optimal crest level

The Delta Committee [12, 13] calculated an acceptable probability of inundation for Central Holland in 1960 of $8{\times}10^{-6}$ per year (Figure 5).

Some approximating calculations performed by Dutch engineers [14] in 2006 indicated an optimal level of 0.2×10^{-3} per year for New Orleans. The city was thought to be protected against a hurricane category 3 with a return period of 30 to 100 years. The present system that was resurrected after Katrina has the planned safety level of 1/100 per year.

The economic criterion presented above should be adopted as a basis for the 'technical' advice to the political decision process. All information of the risk assessment should be available in the political process.

4 THE SAFETY CHAIN

The last years experts in risk management stipulate that prevention of disaster as provided by flood defense systems is inadequate, because the system can fail as shown above. Therefore additional activities have to be undertaken such as the planning and organization of evacuation, the mitigation of damage in case of a disaster, insurance, etc.

In general terms the risk managers advocate the application of the "safety chain" consisting of proaction, prevention, preparation, repression and mitigation, recovery and learning. Proaction means to avoid the danger at all e.g. by not building a city in the Mississippi delta. Prevention indicates the construction of structures that can withstand the force of the rare threat and protect people and goods. Preparation points to planning rescue and mitigation activities in advance. The dictum is: you cannot plan a disaster but the risk management you can. Repression addresses the actual rescue activities after the disaster has struck. Building waterproof facilities or houses on piles that will sustain less damage in case of inundation is indicated by mitigation. Also insuring the properties against the consequences of an inundation falls in this category. Finally the damage should be repaired and the society should be put on it's feet again. This is the recovery phase of risk management.

The risk management experts state that all links of the safety chain have to be addressed by the responsible authorities. This is based on the reasoning that a chain cannot function if an element is omitted. Closer inspection of the safety chain however reveals that it is a parallel system of multiple layers, that is at least as safe as the safest layer. Additionally it should be noted, that that the effectiveness of resources spent in prevention is most probably higher than on repression, because repression becomes only effective after the disaster has occurred and the economic damage is a fact. New Orleans has shown that people and movable property can be saved, but fixed property is subjected to the force of the flood and all economic processes are halted. If the evacuation and recovery expenditure would have been directed at the improvement of the defences, the disaster might have been avoided.

Insurance is a method of repressing the economic damage caused by an uncertain event. The insured pays a insurance premium every year and the

insurer is obliged to refund the main part of the damage if the uncertain event occurs. The insurance premium will be at least equal to the expected value of the loss, the risk. However the insurer must add an allowance for transaction costs, risk aversion and profit. So generally the insurance premium is a factor g higher than the risk. This is especially true if the insured risks are fully dependent, because then all insured are hit simultaneously. This is the case in flood insurance contrary to car or fire insurance.

The model to find the economically optimal risk presented above is easy to adapt for the case of insurance. Let us assume for the sake of simplicity that the insurer covers all damage D in return for a premium that is g times the risk:

$$\text{Premium} = e^{-\frac{(h-A)}{B}} \cdot g \cdot D.$$

Now the total cost of prevention and insurance becomes

$$\text{TC}(h) = I_0 + I_1(h - h_0) + e^{-\frac{(h-A)}{B}} \cdot \frac{g \cdot D}{r}.$$

Applying the same algebra as above the optimal probability of inundation is reduced by a factor g and becomes:

$$p_f^* = e^{-\frac{(h-A)}{B}} = \frac{I_1 B r}{g \cdot D}$$

The conclusion is that the safety of the flood defense should be increased by a factor g and the defenses increased in strength if the damage is privately insured. So for a country like the Netherlands, where a flood will without doubt mean a national disaster, that forces the government to help the stricken people to repair their properties and the infrastructure, an insurance leads to increased cost without clear advantages. If the stricken area is however a small part of a large country, that might be left to it's own devices in recovery, a flood insurance might be wise. Especially if the country's policies lean more towards individual responsibility than state intervention. In the previous part the failure of the insurer was excluded.

It is also interesting to study the optimal division of resources over the elements or layers of a parallel system as the safety chain. To keep it simple the system is limited to two layers with failure probabilities p_1 and p_2. The cost of each system is a linear function of the natural logarithms of the respective failure probabilities p_i (which is similar to the Delta Committee case given above):

$$I(p_1, p_2) = I_0 - I_1 \ln(p_1) - I_2 \ln(p_2).$$

The risk becomes

$$\text{Risk} = p_1 \cdot p_2 \cdot D.$$

The total cost of investment and the present value of the risk equals

$$\text{TC}(p_1, p_2) = I_0 - I_1 \ln(p_1) - I_2 \ln(p_2) + p_1 \cdot p_2 \cdot \frac{D}{r}.$$

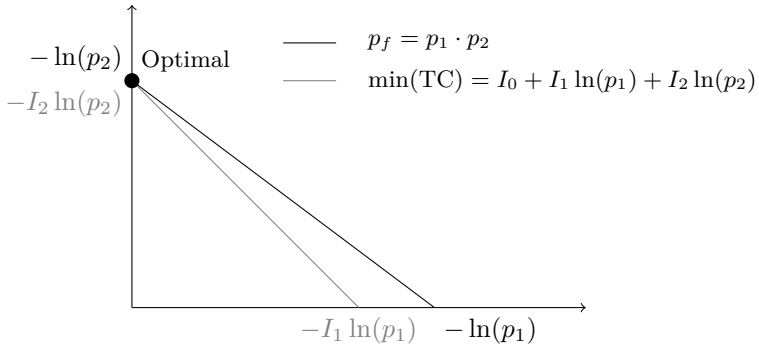


FIGURE 6. The economical optimization of a simple parallel system consisting of two elements

Differentiation with respect to p_i leads to a slightly more complicated result because the minimum lies at the border:

$$p_f^* = \min\left\{\frac{I_1 r}{D}, \frac{I_2 r}{D}\right\}.$$

According to this simple model only the layer with the lowest marginal cost is applied, the other is omitted as shown in Figure 6.

In this simple example the opinion of the risk management experts that all elements of the safety chain must be applied is refuted. Such an example is of course no proof, but it is an indication that the safety chain model in the simple interpretation gives no reliable guidance, because it is a parallel system consisting of layers.

5 CONCLUSIONS

The mathematical risk approach has great advantages compared with the present intuitive. The system of water defenses that is meant to prevent flooding of areas important to mankind, comes at the centre of the analysis that is used as an illustration in this paper. A water defense system is a series-system. The contribution of all elements of the system and of all failure mechanisms of each element to the probability of flooding must be calculated and clearly presented. The safety analysis should include the probability of human failure in the management of the water defense structures is especially attractive and important because human failure is relatively likely. The length effect, meaning that a longer chain is likely to have a weaker link, should be adequately accounted for. The experience in New Orleans has shown, that a long flood defense has indeed weak elements for various reasons. Human error in establishing the chart datum at some locations caused some crest levels to be inadequate. Although some flood

defenses were overtopped without structural failure, others failed more or less immediately when overtopped. At a few locations however the defenses failed without being overtopped. This proved the theoretical prediction that other failure mechanisms contribute to the probability of inundation. Specifically sliding of the 17-th Street Canal flood wall due to soil failure and the undermining of the London Avenue Canal flood wall by piping are vivid illustrations.

An approach was sketched to define the economical optimal level of risk. This was indicated as the acceptable risk. The decision on the level of acceptable risk is a cost/benefit judgement, that must be made from societal point of view. This mathematical optimum should be adopted as a basis for the 'technical' advice to the political decision process. However all information of the risk assessment should be available in the political process. A decision that is political in nature, must be made democratically, because many differing values have to be weighed. The economic optimisation shows however that a fundamental reassessment of the acceptability of the flood risks is justified if the economic activity in the protected areas has grown.

The application of the 'safety chain' consisting of proaction, prevention, preparation, repression/mitigation, recovery and learning was explained and analysed in some depth. It was observed that effectiveness of resources spent in prevention is most probably higher than on repression, because repression becomes only effective after the disaster has occurred and at least the economic damage is a fact.

As an example of repression insurance was analysed. It appeared that insurance forces to a higher level of protection because the insurance premium exceeds the risk by some factor. The total cost of prevention and private insurance will increase compared to a state insurance. So in countries like the Netherlands where a flood will be an national disaster, insuring flood damage seems ill advised. A community that cannot count on national aid in case of a disaster might be wise to opt for insurance.

Finally a parallel system of two layers was economically optimized under the assumption that any level of safety could be reached at a cost that is a linear function of the logarithm of the failure probability. It appeared that the optimal investment was limited to one layer of protection, the layer with the lower marginal cost. This refutes in some sense the quick conclusion of the simple safety chain reasoning that every element should be addressed.

It is clear from the examples in this paper that the mathematical methods of risk analysis and probabilistic reasoning are great aids in the design and the understanding modern safety systems.

Bibliography

[1] *Probabilistic design of sea defences.* Technical Advisory Committee on Water Retaining Structures, CUR, Gouda, 1989.

[2] M. Hohenbichler and R. Rackwitz. First-order concepts in system reliability. *Structural Safety*, 1:177–188, 1983.

[3] A. S. Nowak and K. R. Collins. *Reliability of structures*. McGraw-Hill, 2000.

[4] Mark G. Stewart and Robert E. Melchers. *Probabilistic risk assessment of engineering systems*. Chapman and Hall, London, 1997.

[5] C. Guedes Soares. Dealing with strength degradation in structural reliability. In P. van Gelder, A. Roos, and H. Vrijling, editors, *In Risk-Based Design of Civil Structures*, number 01-1, pages 31–49. Communications on hydraulic and geotechnical engineering, 2001. ISSN 0169-6548.

[6] J. K. Vrijling. Review of "Dealing with strength degradation in structural reliability". In Pieter van Gelder, Alex Roos, and Han Vrijling, editors, *In Risk-Based Design of Civil Structures*, number 01-1, pages 51–53. Communications on hydraulic and geotechnical engineering, 2001. ISSN 0169-6548.

[7] P. H. A. J. M. van Gelder. *Statistical methods for the risk-based design of civil structures*. PhD thesis, Delft University of Technology, 1999.

[8] ASCE. *The New Orleans Hurricane protection System, What went wrong and why*. American Society of Civil Engineers, Reston, Virginia, 2007.

[9] D. Blockley. *Engineering safety*. McGraw-Hill, London, 1992.

[10] J. K. Vrijling, W. Van Hengel, and R. J. Houben. Acceptable risk as a basis for design. *Reliability Engineering and System Safety*, 59(1):141–150, 1998.

[11] J. K. Vrijling and P. H. A. J. M. Van Gelder. An analysis of the valuation of a human life. In M. P. Cottam, D. W. Harvey, R. P. Pape, and Tait J., editors, *Foresight and Precaution: Proceedings of ESREL 2000, SARS and SRA-Europe Annual Conference, Edinburgh, Scotland, May 15-17, 2000*, volume 1, pages 197–200, Lisse, the Netherlands, 2000. Balkema.

[12] D. van Dantzig. Economic decision problems for flood prevention. *Econometrica*, 24:276–287, 1956.

[13] D. Van Dantzig and J. Kriens. The economic decision problem of safeguarding the netherlands against floods. Technical report, Report of Delta Committee, Part 3, Section II.2 (in Dutch), The Hague, 1960.

[14] J. Dijkman, editor. *A Dutch perspective on Coastal Louisiana: Flood risk reduction and landscape stabilization*. Number WL-Z4307. Netherlands Water Partnership (NWP), Delft, the Netherlands, October 2007.

This page intentionally left blank

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Relative material loss: a maintenance inspection methodology for approximating material loss on in-service marine structures

Robert A. Ernsting* – Northrop Grumman Shipbuilding, Newport News, Virginia, U.S.A., Thomas A. Mazzuchi and Shahram Sarkani – The George Washington University, Washington D.C., USA

Abstract. This paper describes a new maintenance inspection methodology called *relative material loss* (RML) used for approximating the material loss contribution on each plate side separating two or more dissimilar marine environments. The new methodology leverages actual "at sea" environmental and operational conditions by defining relationships between the dissimilar environments and solving for the material loss on each plate side. The RML theory and a case study using a sixty five year old in-service structure; a dry dock caisson gate is presented.

1 INTRODUCTION

In 2009, the American Society of Civil Engineers (ASCE), estimated that the United States must invest $2.2 Trillion over five years to refurbish its crumbling infrastructure. This is up from $1.6 Trillion reported in 2005 [1]. To offset the staggering rising costs, infrastructure researchers are developing maintenance optimization methodologies to support the growing demand for service life extensions over expensive replacement strategies.

For example, the Dutch polders in the Netherlands rely on the safe and reliable performance of its 2500 km of dykes, dams and barriers to protect its citizens and low-lying cities from the North Sea. However, deterioration mechanisms such as dyke settlement, subsoil consolidation, and relative sea-level rise create a constant engineering challenge in order to maintain these complex civil infrastructures [2]. Through the use of probabilistic modeling, Dutch civil engineers have created optimization maintenance methodologies for the inspection and predicted maintenance of these critical national structures [3]. For example, condition-based maintenance (CBM) models utilize Gamma processes for modeling asset deterioration while Poisson processes are employed to model service load events [4]. Degradation of the asset occurs stochastically over time to a predetermined service level

*corresponding author: Northrop Grumman Shipbuilding, 4101 Washington Avenue, Newport News, VA 23607, U.S.A.; telephone: +1-(757) 688 7469, e-mail: ernsting@gwu.edu

at which a maintenance action is required. Non-invasive inspections occur at predetermined time intervals to determine remaining service life and allow ample time for maintenance planning and funding. The Dutch have used the Gamma deterioration models to model paint deterioration on steel structures, establish dyke heightening strategies, optimize sand nourishment strategies, and predict severe long shore rock transport along berm breakwaters [2, 5, 6, 7].

2 RELATIVE MATERIAL LOSS

It has been suggested that marine corrosion modeling research lacks a framework for analyzing material loss data and applying probabilistic corrosion models [8]. Some research is conducted in laboratory environments under controlled conditions [9], while others are conducted by taking actual thickness measurements directly from in-service structures such as cargo tankers [10]. This paper proposes a new maintenance inspection methodology that presents a new paradigm for researchers to apply probabilistic prediction models. The methodology is based upon a new theory called *relative material loss*, or RML [11, 12].

2.1 Definitions

Before describing the RML theory, the following definitions are presented:

Relative Material Loss – A maintenance inspection methodology for approximating material loss contribution on each side of structural shell plating subjected to dissimilar marine environments [11].

Material Loss Contribution – The amount of material loss on a structural member that is attributable to the environment from which it exists. Material loss contribution is designed by c_e, where e corresponds to the environment causing the material loss.

Relative Loss Equations – Mathematical relationships or equations defined across various environmental boundaries (such as shell plating or sheet piling) and solved simultaneously to suggest solutions.

Laterally homogeneous – Environmental parameters causing material loss are equivalent laterally on either side of a point within the environment.

Longitudinally heterogeneous – Environmental parameters causing material loss are not equivalent longitudinally above or below a point within an environment.

Environment – An environment wholly exists if a laterally homogeneous and longitudinally heterogeneous condition exists.

2.2 RML Theory

Currently, ultrasonic measuring equipment can only determine *total* remaining material across structural shell plating; the device cannot distinguish the

amount of material loss contribution on each plate side. Relative material loss (RML) theory leverages actual in-service environmental and operational conditions and, by establishing relationships between them, suggests solutions for otherwise indeterminate material loss variables. In much the same manner as structural engineers use free body diagrams to isolate joints on a truss to determine member forces, relative loss (RL) equations are defined across various environmental boundaries (i.e. shell plating) and solved simultaneously to suggest solutions (or material loss contributions). In the next section, the RML theory is developed using a series of cases that build upon each other.

In the case where a single steel plate is immersed in a single, homogeneous environment, the calculation of the amount of material loss on each plate side is straightforward. It is the measured total plate thickness loss (or wastage) divided by 2. In the case where a single plate is subjected to two dissimilar environments, such as with the shell plating of an above-ground tank, the solution is indeterminate due to having two unknown variables such that:

$$c_A + c_B = W, \tag{1}$$

where c_A and c_B are the material losses on each side of a single plate subjected to environments A and B and W is the measured total wastage across the plate. However, a solution is possible if a second independent equation in terms of unknown variables c_A and c_B can be defined, assumed or calculated by other means. In the above-ground tank example, a reasonable assumption is made that the exterior side receives periodic maintenance to the extent that the material loss contribution is approximately zero compared to its interior side, providing the second equation needed for the solution. Therefore, an assumption is made that $c_B \approx 0$. Other means to approximate c_B include using deterministic formulas or probabilistic models found in literature.

Figure 1 provides examples of structural systems exposed to three dissimilar environments, A, B, and C. The cellular cofferdam retaining wall (Figure 1a) is comprised of steel sheet piles. The cofferdam environments consist of clay and gravel soil conditions under different hydrostatic conditions and a marine atmospheric (sea air). The double hull tanker shell plating (Figure 1b) is exposed to crude oil, confined ballast tank atmospheric and seawater environments. The dry dock caisson gate shell plating (Figure 1c) is exposed to brackish river water, ballast tank water and marine atmospheric conditions.

The model for these examples is presented (see Figure 2). Two steel plates 1 and 2 are exposed to three dissimilar environments A, B, and C. Each steel plate has an original plate thickness, d_0 and, over time, t, experience material thickness losses, c_A, c_B and c_C, caused by their environments. In the field, inspectors use ultrasonic devices to measure remaining plate thicknesses, d_1 and d_2, on plates 1 and 2, respectively at time, t. Total

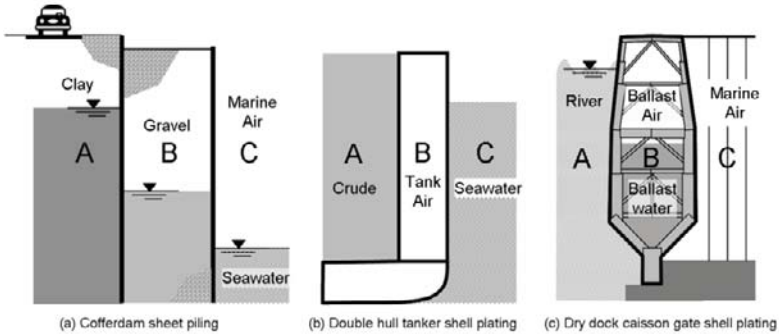


FIGURE 1. Examples of structures that are exposed to dissimilar environments, A, B and C

wastages, W_1 and W_2 at time, t, are calculated:

$$W_1(t) = d_{0_1} - d_1(t), \tag{2}$$
$$W_2(t) = d_{0_2} - d_2(t). \tag{3}$$

The original plate thicknesses d_0 is based on *actual* and not *nominal* thickness. However, the actual plate thickness is not often known. Therefore, using the nominal plate thickness can frequently produce "negative" wastage numbers, such that $W < 0$. If this occurs, then the plate should be checked against its allowable manufacturers mill overage tolerance and d_0 calibrated accordingly. For example, using ASTM A6 [13].

Referring to Figure 2, material loss contributions $c_A(t)$ and $c_B(t)$ are expressed in terms of known wastages, $W_1(t)$ and $W_2(t)$ and unknown material loss contributions $c_B(t)$ and $c_C(t)$ such that:

$$c_A(t) = W_1(t) - c_B(t), \tag{4}$$
$$c_B(t) = W_2(t) - c_C(t). \tag{5}$$

Due to high variability inherent with material loss data, it is appropriate to represent material loss probabilistically in the form of a material loss function [14]. Melchers [15] and Qin and Cui [16] utilize a generic material loss function as a framework for calculating probabilistic material loss, $c(t, \boldsymbol{P}, \boldsymbol{E})$, as a function of time, t:

$$c(t, \boldsymbol{P}, \boldsymbol{E}) = b(t, \boldsymbol{P}, \boldsymbol{E}) \times f(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon(t, \boldsymbol{P}, \boldsymbol{E}), \tag{6}$$

where $f(t, \boldsymbol{P}, \boldsymbol{E})$ is the mean-value corrosion loss function; $b(t, \boldsymbol{P}, \boldsymbol{E})$ is a bias function; $\epsilon(t, \boldsymbol{P}, \boldsymbol{E})$ is a zero-mean uncertainty function such that $e(t, \boldsymbol{P}, \boldsymbol{E}) \approx N(0, s)$; E is a vector of environmental parameters that influence corrosion such as water temperature, dissolved oxygen, pH, pollutants,

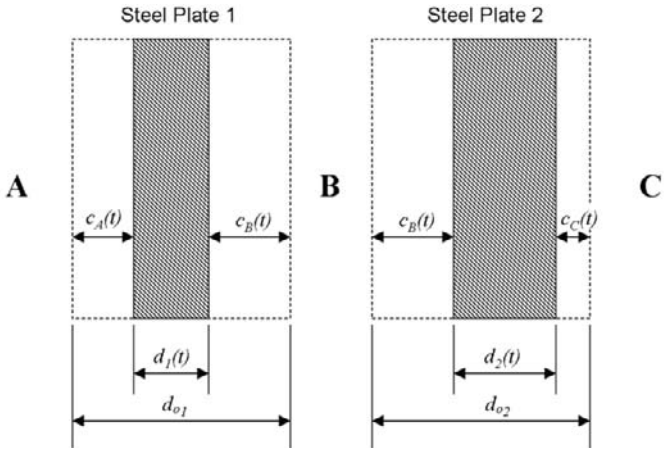


FIGURE 2. Steel plates 1 and 2 of original thicknesses d_{0_1} and d_{0_2} separating three dissimilar environments, A, B and C and influencing material losses c_A, c_B and c_C over time, t

wave action, water velocity [17]. When the function is used on in-service structures, a vector, P, is added to represent parameters that resist material loss such as coatings and cathodic protection systems [16]. There are numerous time-based corrosion prediction models found in the literature that can be used to define $f(t, \boldsymbol{P}, \boldsymbol{E})$ [16, 17, 18, 19, 20].

The bias function, $b(t, \boldsymbol{P}, \boldsymbol{E})$, is multiplicative rather than additive for it is used to calibrate $f(t, \boldsymbol{P}, \boldsymbol{E})$ with $c(t, \boldsymbol{P}, \boldsymbol{E})$. When an accurate model for $f(t, \boldsymbol{P}, \boldsymbol{E})$ exists, it exactly represents $c(t, \boldsymbol{P}, \boldsymbol{E})$ and the bias function, $b(t, \boldsymbol{P}, \boldsymbol{E})$ is defined as unity [21]. Along with a well calibrated model, careful inspection and sound sampling methodologies must be employed. Assuming this condition where bias is unity, Eq. (6) is simplified:

$$c(t, \boldsymbol{P}, \boldsymbol{E}) = f(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon(t, \boldsymbol{P}, \boldsymbol{E}). \tag{7}$$

2.3 Relative Loss Equations

With respect to Figure 2, Eqs. (4) and (5) are used to define the mean-value corrosion loss functions, $f(t, \boldsymbol{P}, \boldsymbol{E})$ used for this study. Therefore, the following is given:

$$f_A(t, \boldsymbol{P}, \boldsymbol{E}) = W_1(t) - c_B(t, \boldsymbol{P}, \boldsymbol{E}), \tag{8}$$
$$f_B(t, \boldsymbol{P}, \boldsymbol{E}) = W_1(t) - c_C(t, \boldsymbol{P}, \boldsymbol{E}). \tag{9}$$

Through substitution, the right-hand side of Eqs. (8) and (9) replace $f(t, \boldsymbol{P}, \boldsymbol{E})$ in Eq. (7) to generate two relative loss (RL) equations:

$$c_A(t, \boldsymbol{P}, \boldsymbol{E}) = W_1(t) - c_B(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon_B(t, \boldsymbol{P}, \boldsymbol{E}), \tag{10}$$

$$c_B(t, \boldsymbol{P}, \boldsymbol{E}) = W_1(t) - c_C(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon_C(t, \boldsymbol{P}, \boldsymbol{E}). \tag{11}$$

Subtracting Eqs. (10) and (11) and solving for $c_A(t, \boldsymbol{P}, \boldsymbol{E})$ yields:

$$c_A(t, \boldsymbol{P}, \boldsymbol{E}) = W_1(t) - W_2(t) + c_C(t, \boldsymbol{P}, \boldsymbol{E}) - \epsilon_B(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon_C(t, \boldsymbol{P}, \boldsymbol{E}). \tag{12}$$

Notice that the material loss contribution, $c_B(t, \boldsymbol{P}, \boldsymbol{E})$ is irrelevant. The relationships in Eqs. (2) and (3) are substituted for $W_1(t)$ and $W_2(t)$ in Eq. (12) to account for the original plate thicknesses d_{0_1} and d_{0_2}, and remaining plate thicknesses $d_1(t)$ and $d_2(t)$:

$$c_A(t, \boldsymbol{P}, \boldsymbol{E}) = d_{0_1} - d_{0_2} + d_2(t) - d_1(t) - c_C(t, \boldsymbol{P}, \boldsymbol{E}) - \epsilon_B(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon_C(t, \boldsymbol{P}, \boldsymbol{E}). \tag{13}$$

Notice in cases where $d_{0_1} = d_{0_2}$, original plate thickness (and associated actual mill tolerance) provides no additional information. Under this condition, the RL equation simplifies:

$$c_A(t, \boldsymbol{P}, \boldsymbol{E}) = d_2(t) - d_1(t) - c_C(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon_B(t, \boldsymbol{P}, \boldsymbol{E}) - \epsilon_C(t, \boldsymbol{P}, \boldsymbol{E}). \tag{14}$$

The relationship between the two zero-mean uncertainty functions $\epsilon_B(t, \boldsymbol{P}, \boldsymbol{E})$ and $\epsilon_C(t, \boldsymbol{P}, \boldsymbol{E})$ suggests that uncertainty is reduced if $\epsilon_B(t, \boldsymbol{P}, \boldsymbol{E}) \approx \epsilon_C(t, \boldsymbol{P}, \boldsymbol{E})$. However, this suggestion violates the rules of second moment algebra for variance [22]. Further research is warranted to explore the uncertainty reduction suggestion within the context of relative material loss theory since various modeling techniques can be employed. For the purpose of this study, the uncertainty term will be designated as $\epsilon'(t, \boldsymbol{P}, \boldsymbol{E})$ to indicate that uncertainty has changed and an assumption is made that $\epsilon'(t, \boldsymbol{P}, \boldsymbol{E})$ is near zero.

$$c_A(t, \boldsymbol{P}, \boldsymbol{E}) = d_2(t) - d_1(t) + c_C(t, \boldsymbol{P}, \boldsymbol{E}) + \epsilon'(t, \boldsymbol{P}, \boldsymbol{E}). \tag{15}$$

Eq. (15) is a specific relative loss (RL) equation for situations where two steel plates are separating three dissimilar environments (as in the Figure 1 examples) and have the same original plate thickness. Simply stated, the material loss contribution, $c_A(t, \boldsymbol{P}, \boldsymbol{E})$, is a function of the measured plate thicknesses $d_1(t)$ and $d_2(t)$ of the two shell plates plus the material loss contribution, $c_C(t, \boldsymbol{P}, \boldsymbol{E})$ plus all applicable uncertainty, $\epsilon'(t, \boldsymbol{P}, \boldsymbol{E})$.

It is important to note that a relationship exists between the two external (or opposing) sides plate surfaces are expressed in terms of the two measured total plate thicknesses. This RL equation can be used either deterministically by plugging field data directly into the RL equation or probabilistically using the corrosion prediction models found in the literature. If $c_c(t)$ cannot be reasonably assumed, a value for $c_C(t)$ can be estimated from empirical formulas found in literature or by defining another independent RL equation.

2.4 Laterally homogeneity and longitudinally heterogeneity

ASSUMPTION: the marine immersion environments are stratified such that the conditions that cause material loss are laterally homogeneous and longitudinally heterogeneous (as per Section 2.1).

The assumption implies that in order for a single environment to exist, it must be laterally homogeneous and longitudinally heterogeneous. To test for laterally homogeneity, a one-way analysis of variance (ANOVA) test is performed on a collection of thickness measurements laterally subdivided equally into two groups. Set the null hypothesis, as the two means of the subdivided groups are equal and the alternative hypothesis, as the two means are not equal. For the environments to be laterally homogeneous, the null hypothesis must fail to reject. If the null hypothesis does not fail to reject, then explore the possibility that another dissimilar environment exists by repeating the test at various locations laterally along the structure.

Once lateral homogeneity is established, a test is performed to verify longitudinal heterogeneity. To test for this, a one-way ANOVA test is performed on a collection of thickness measurements longitudinally subdivided into groups at each level of the structure. Set the null hypothesis as all means of the subdivided groups at each level are equal and the alternative hypothesis to at least one level is not equal. For the environments to be longitudinally heterogeneous, at least one level will be significantly different, causing the null hypothesis to reject [11].

3 DRY DOCK CAISSON GATE

A dry dock caisson gate is a floatable steel vessel that is submerged at the free water end of a dry dock to seal the dock from the river (Figure 1c). As the dock is dewatered, hydrostatic forces from the riverside build vertically along the caisson gate, pressing the gate against the concrete abutment of the dry dock to form a watertight seal. The dry dock caisson gate has recently been suggested as a unique new research platform for studying in-service marine corrosion. This is due to its (1) controlled and predictable marine environments and (2) recent increase in the regulatory inspections [12].

3.1 Northrop Grumman Shipbuilding caisson gate

A caisson gate that services a dry dock at Northrop Grumman Shipbuilding located near the mouth of the James River in Newport News, Virginia, USA was used for RML theory validation. Shell plating thickness readings from a recent NAVSEA inspection provide the platform for testing and validating the relative material loss theory. At the time that the shell plating thickness readings were taken in 2005, the caisson gate was 65 years old. The caisson gate was routinely dry docked and overhauled approximately once every 8-10 years.

Ultrasonic pulse echo measurements were taken on the river and marine

air shell plating of the caisson gate on a 2.4m grid pattern. Tests for laterally homogeneity and longitudinally heterogeneity were performed as described in section 2.4. The caisson gate environments are laterally homogeneous for the null hypothesis failed to reject ($\alpha = 0.05$). River Side: p-value = 0.421, Marine Air Side: p-value = 0.180. The caisson gate environments are longitudinally heterogeneous for at least one level is significantly different, causing the null hypothesis to reject ($\alpha = 0.05$). River Side: p-value $<$ 0.001, Marine Air Side: p-value $<$ 0.001.

To describe existing material loss conditions deterministically and account for variability (i.e. 95% confidence intervals), a least squares regression technique [23] is chosen over a spline technique. Using regression software, ten polynomial regression models (PRM) are created at each caisson gate level (A, B, C, D and E) and on each side (River and Marine Air) of the caisson gate (see positions indicated in the lower right-hand corner of Figure 3). The order of the polynomial varied between models and was selected based on the lowest p-value and maximum R-square (adjusted) values obtainable through trial and error. Table 1 summarizes the R-squared and p-values of the ten polynomial regression models.

PRM	R^2	Adjusted R^2	ANOVA p-value
River-A	38.5	20.1	0.166
River-B	59.8	56.4	0.001
River-C	no correlation, use mean		
River-D	54.7	34.6	0.098
River-E	70.8	52.5	0.044
Air-A	89.8	77.8	0.013
Air-B	57.3	38.7	0.076
Air-C	46.0	36.2	0.034
Air-D	56.1	42.9	0.035
Air-E	77.6	67.6	0.005

TABLE 1. Polynomial Regression Models: R-Squared and p-values

The PRM's are used to calculate the remaining plate thicknesses, d at any given caisson gate frame location number, x. As an example, Figure 3 depicts the PRM for River Side, Level "D" [designated herein as $f_{\mathrm{River},D}(x)$]. The thick solid line represents the fitted point estimate and the dashed line represent the 95% upper and lower confidence intervals. To provide an independent validation of the PRM model, a locally weighted estimated scatter smoothing technique (LOWESS) is superimposed [24]. The goodness-of-fit of the LOWESS models is performed by visually checking the residuals for normality, autocorrelation and heteroscedasticity.

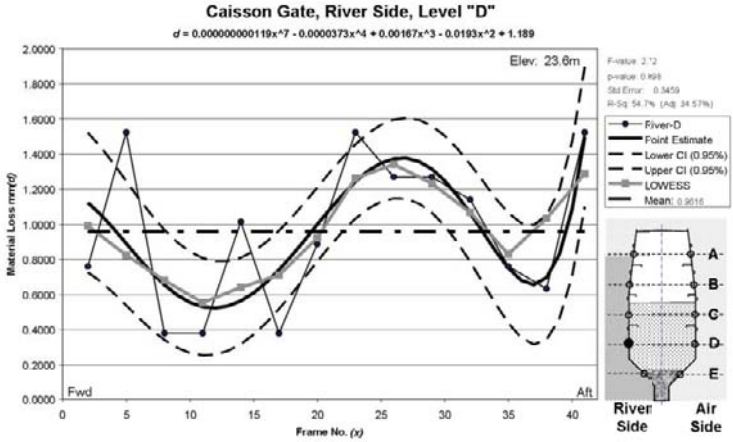


FIGURE 3. Polynomial regression model using least squares

3.2 Relative Loss Profiles

Using relative loss Eq. (15), five RL equations are constructed substituting the PRM functions for d_1 and d_2 at each caisson gate level. PRM solutions for c_{River} at each of the five levels are calculated for all x.

$$c_{\text{River}} = f_{\text{Air},n}(x) - f_{\text{River},n}(x) + c_{\text{Air}}, \tag{16}$$

where:

$f_{s,n}(x)$ = polynomial regression models for either river side or marine air side, such that
s = "Air" or "River" side
n = level A, B, C, D or E
x = caisson gate frame location number (2, 3, 4, ..., x, ..., 40, 41)
c_{River} = Relative material loss contribution of river side exterior plating
c_{Air} = Relative material loss contribution of air side exterior plating (assumption: $c_{\text{Air}} \approx 0$)

Since a deterministic approach was chosen for this case study, t, $\boldsymbol{P}$, and $\boldsymbol{E}$ are dropped from the RL equations. Also, since the exterior side of the caisson gate is exposed to marine atmospheric conditions and assumed routinely maintained, an assumption is made that $c_{\text{Air}} \approx 0$. c_B (inside ballast tank) is calculated by subtracting c_{Air} from the total measured material loss, W_2 (Eq. 3). The five RL equations are solved at each frame number

location, x along the caisson gate shell plating from frame number location 2 to 41 and relative loss profiles created.

4 DISCUSSION

Figure 4 provides the total material loss profile along the riverside as measured in the field. The thick black line represents the mean value profile of total material loss measured. The thin gray lines are vertical profiles at each caisson gate frame number location, x and provide a "bootstrap" approximate confidence interval of the point estimate. Note: since Figure 4 is only for illustrating total material loss vertically along the shell plating and is not used in the RML equations, a simple piecewise linear function was chosen.

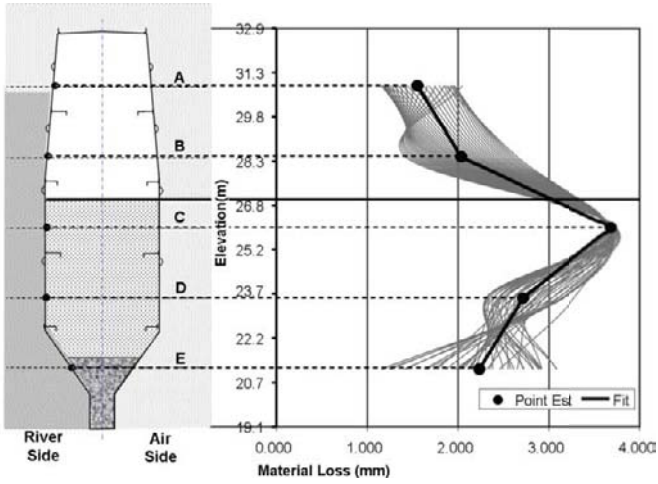


FIGURE 4. Total material loss profile: river side

Figure 5 provides the relative material loss profile of the riverside exterior shell plating. Again, the thin gray lines are vertical profiles at each caisson gate frame number location, x and provide a "bootstrap" approximate confidence interval of the point estimate. Note again that inflection points of the thin gray lines appear to correspond with anomalies with the structure - change in shell plating orientation at elevation 22.2, horizontal plate stiffeners at elevations 22.5 and 25.2, and top of ballast water at elevation 26.8. This suggests that RML could potentially be used to locate structural anomalies hidden within structures and tanks.

The mean RML at each level is calculated and represented on Figure 5. A triangle indicates the mean RML value for marine atmospheric at level A. This correlates to the assumption made earlier that the contribution due to marine atmospheric conditions is near zero. Fours dots indicate the

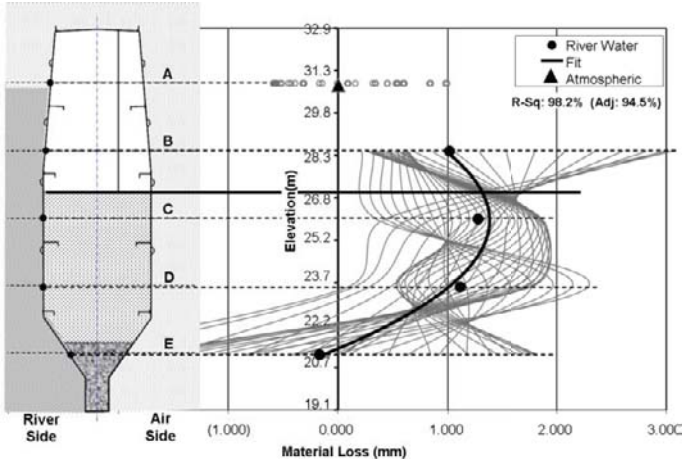


FIGURE 5. Relative material loss profile: river side, exterior shell plating

mean RML values of the material loss contribution due to marine seawater immersion at levels B, C, D, and E. Using a least squares technique, a fitted curve is drawn between the four points with an R-Sq adjusted = 94.5%. The curve has been shown to have a high correlation with the water dissolved oxygen content (r-value=.925) and water temperature (r-value=.768). Also, note the mean RML value at level E is near zero.

Figure 6 provides the relative material loss profile of the interior shell plating. The ballast tank atmospheric condition is conducive to high humidity and high temperature. The higher mean RML value at level A compared to level B is contributable to (1) higher surface temperatures at level A due to radiant heat from direct sun light and (2) increased time of wetness (TOW) due to dew build up on the inside surface from fluctuating day-time and night-time temperatures [25]. Comparing mean RML value at level A in Figure 6 with the mean RML value at level A in Figure 5, it is suggested that the majority of the material loss is occurring on the inside of the structure. The ballast water trend between levels C and D in Figure 6 correlate well with water dissolved oxygen content (r-value=.918) and water temperature (r-value=.987).

Although the variability at level E is high, the mean RML value indicates high material loss at this region. This is plausible for the concrete placed in the region is original and 65 years old. The concrete as it ages will tend to crack and develop fissures over its long life span and expose the inside shell plating surface to salts and chlorides from the seawater ballast directly above. Since the inspection of the inside, shell plate surface at level E is difficult, RML proves to be a valuable methodology for future monitoring in this region.

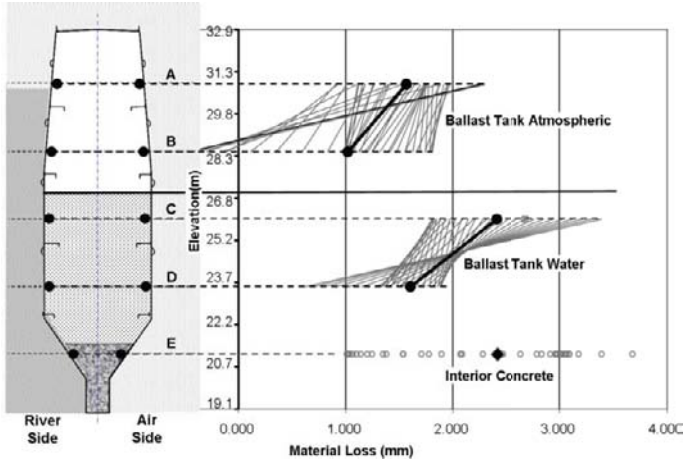


FIGURE 6. Relative material loss profile: interior shell plating

The example given is specific to situations where two steel plates separate three dissimilar environments such as the cellular cofferdam retaining wall, double hull tanker shell plating and dry dock caisson gates in Figure 1. From a general perspective, relative material loss theory can be applied to any situation where n structures isolate $n+1$ dissimilar environments. However, as n becomes large, additional RL relationships are needed to resolve the indeterminacy problems.

5 CONCLUSION

This paper proposes a paradigm shift of corrosion and material loss research and introduces a new maintenance inspection methodology called *relative material loss* (or RML). Material loss contribution on each side of a plate separating dissimilar marine environments is approximated by establishing mathematical relationships between dissimilar environments using relative loss (RL) equations and solving the equations simultaneously. RML can be applied either (1) deterministically at any time to locate areas of unusual degradation or (2) probabilistically using any of the time-based corrosion prediction models found in the literature. The methodology was demonstrated on a 65-year old marine structure; a dry dock caisson gate where material loss profiles were created and shown to correlate closely with water temperature and dissolved oxygen content. As the caisson gate is periodically inspected, new information can be introduced and the RML profiles updated to erflect degradation as a function of time. Furthermore, by mapping material loss to environmental and operational parameters over extended periods of time, improved corrosion prediction models for in-service

structures are possible. The potential use of RML on in-service structures is far reaching and has potential applications on dry dock caisson gates, ballast tanks, ship hull structures, bridge abutments, sheet pile cofferdams, underground piping systems, storage tanks, offshore oil platforms and flood control/sluice gates.

Acknowledgments

The author would like to acknowledge Northrop Grumman Shipbuilding-Newport News for providing access to the material loss data pertinent for this research. Additional acknowledgment to Dr. Thomas Mazzuchi, Shahram Sarkani, Harvey Hack, Frank Allario and Robert Melchers for providing advice and guidance. And lastly, this paper is dedicated to the late professor Jan M. van Noortwijk whom provided the inspiration as well as anchor paper that spawned this research.

Bibliography

[1] ASCE report card for America's infrastructure. Technical report, ASCE, 2009. URL `http://www.infrastructurereportcard.org/`.

[2] L. J. P. Speijker, J. M. van Noortwijk, M. Kok, and R. M. Cooke. Optimal maintenance decisions for dikes. *Probability in the Engineering and Informational Sciences*, 14(4):101–121, 2000.

[3] D. M. Frangopol, M. J. Kallen, and J. M. van Noortwijk. Probabilistic models for life-cycle performance of deteriorating structures: review and future directions. *Progress in Structural Engineering and Materials*, 6(4):197–212, 2004.

[4] J. M. van Noortwijk, J. A. M. van der Weide, M. J. Kallen, and M. D. Pandey. Gamma processes and peaks-over-threshold distributions for time-dependent reliability. *Reliability Engineering and System Safety*, 92(12): 1651–1658, 2007.

[5] R. P. Nicolai, R. Dekker, and J. M. van Noortwijk. A comparison of models for measurable deterioration: an application to coatings on steel structures. *Reliability Engineering and System Safety*, 92(12):1635–1650, 2007.

[6] J. M. van Noortwijk and E. B. Peerbolte. Optimal sand nourishment decisions. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 126(1): 30–38, 2000.

[7] J. M. van Noortwijk and P. H. A. J. M. van Gelder. Optimal maintenance decisions for berm breakwaters. *Structural Safety*, 18(4):293–309, 1996.

[8] R. E. Melchers. Probabilistic models for corrosion in structural reliability assessment – part 1: Empirical models. *Journal of Offshore Mechanics and Arctic Engineering*, 125(4):264–271, 2003.

[9] R. E. Melchers. Probabilistic modelling of marine corrosion of steel specimens. In *95: 5th International Offshore & Polar Engineering Conference*, pages 204–210, The Hague, 1995. ISOPE: International Society of Offshore and Polar Engineers.

[10] Y. Garbatov, C. Guedes Soares, and G. Wang. Nonlinear time dependent corrosion wastage of deck plates of ballast and cargo tanks of tankers. *Journal*

of Offshore Mechanics and Arctic Engineering, 129:48, 2007.

[11] R. A. Ernsting. *Methodology for Approximating Material Loss on Structural Plating Subjected To Dissimilar Marine Environments*. PhD thesis, The George Washington University, Washington D.C., 2009.

[12] R. A. Ernsting, T. A. Mazzuchi, and S. Sarkani. Using relative material loss to evaluate a dry dock caisson gate. *Materials Performance*, 48:5, 2009.

[13] ASTM. Standard specification for general requirements for rolled structural steel bars, plates, shapes, and sheet piling. Technical Report ASTM A6/A6M-07, American Society for Testing and Materials, West Conshohocken, 2007.

[14] G. Wang, J. Spencer, and H. Sun. Assessment of corrosion risks to aging ships using an experience database. *Journal of Offshore Mechanics and Arctic Engineering*, 127(2):167–174, 2005.

[15] R. E. Melchers. Modeling of marine immersion corrosion for mild and low-alloy steels part 2: Uncertainty estimation. *Corrosion(USA)*, 59(4):335–344, 2003.

[16] S. Qin and W. Cui. Effect of corrosion models on the time-dependent reliability of steel plated elements. *Marine Structures*, 16(1):15–34, 2003.

[17] R. E. Melchers. Recent progress in the modeling of corrosion of structural steel immersed in seawaters. *Journal of Infrastructure Systems*, 12(3):154–162, 2006.

[18] C. Guedes Soares, Y. Garbatov, A. Zayed, G. Wang, R. E. Melchers, J. K. Paik, and W. Cui. Non-linear corrosion model for immersed steel plates accounting for environmental factors. *Transactions SNAME*, 115(1):19–21, 2005.

[19] J. K. Paik, A. K. Thayamballi, Y. I. Park, and J. S. Hwang. A time-dependent corrosion wastage model for seawater ballast tank structures of ships. *Corrosion Science*, 46(2):471–486, 2004.

[20] G. Wang, J. Spencer, and T. Elsayed. Estimation of corrosion rates of structural members in oil tankers. In *OMAE 2003, 22nd International Conference on Offshore Mechanics and Arctic Engineering*, Cancun, Mexico, 2003. ASME.

[21] R. E. Melchers. Probabilistic model for marine corrosion of steel for structural reliability assessment. *Journal of Structural Engineering*, 129(11):1484–1493, 2003.

[22] J. R. Benjamin and C. A. Cornell. *Probability, statistics, and decision for civil engineers*. McGraw-Hill, New York, 1970.

[23] H. J. Motulsky and L. A. Ransnas. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. *The FASEB Journal*, 1 (5):365–374, 1987.

[24] W. S. Cleveland and S. J. Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[25] F. Brennan, C. A. Carlsen, C. Daley, Y. Garbatov, L. Ivanov, C. M. Rizzo, B. C. Simonsen, N. Yamamoto, and H. Z. Zhuang. Issc committee v.6: Condition assessment of aged ships. *16th International Ship and Offshore Structures Congress*, 2:265–315, 2006.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Nonparametric predictive system reliability with all subsystems consisting of one type of component

Frank P.A. Coolen*, Ahmad M. Aboalkhair and Iain M. MacPhee
– Durham University, Durham, United Kingdom

Abstract. Recently we have presented nonparametric predictive inference (NPI) for system reliability [1, 2], with specific attention to redundancy allocation. Series systems were considered in which each subsystem i is a k_i-out-of-m_i system. The different subsystems were assumed to consist of different types of components, each type having undergone prior success-failure testing. This work uses NPI for Bernoulli variables [3], which enables prediction for m future variables based on n observations, without the need of a prior distribution. In this paper, we present a generalization of these results by considering multiple subsystems which all consist of one type of component, which provides an important step to wider applicability of this approach.

1 INTRODUCTION

During recent decades, generalization of the standard theory of probability, in which a single value is used to quantify uncertainty for a specific event, by the use of lower and upper probabilities has become increasingly popular, see [4] for an introductory overview from the perspective of reliability theory and applications. The main idea is that, for an event A, a lower probability $\underline{P}(A)$ and upper probability $\overline{P}(A)$ are specified, such that $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, with classical precise probability appearing in the special case with $\underline{P}(A) = \overline{P}(A)$. Like precise probability, lower and upper probabilities have different possible interpretations, including a subjective interpretation in terms of buying prices for gambles. Informally, a lower probability $\underline{P}(A)$ can be interpreted as reflecting the evidence in support of event A, which makes focus on lower probability for system functioning natural and attractive in reliability studies, we use this as the reliability measure of interest throughout this paper. For completeness, however, we also present the corresponding upper probability $\overline{P}(A)$, which can be interpreted by considering that $1 - \overline{P}(A)$ reflects the evidence against event A, so in support of the complementary event A^c. The lower and upper probabilities presented in this paper are naturally linked by the conjugacy

*corresponding author: Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, England; telephone: +44-(0)191 334 3048, fax: +44-(0)191 334 3051, e-mail: frank.coolen@durham.ac.uk

property $\overline{P}(A) = 1 - \underline{P}(A^c)$ [5, 3].

For this paper it suffices to regard the lower and upper probabilities as the optimal bounds for a probability that can be derived from limited assumptions, indeed a major benefit of lower and upper probabilities for statistical inference is that one does not need to make modelling assumptions that are strong enough to derive precise probabilities. For the approach presented in this paper, the main benefit is that predictive inference is possible without the need to assume a prior probability distribution, as is the case in Bayesian statistics.

Coolen [3] presented lower and upper probabilities for prediction of Bernoulli random quantities, which have strong consistency properties within the theory of interval probability [5]. These lower and upper probabilities followed from an assumed underlying latent variable model, with future outcomes of random quantities related to data by Hill's assumption $A_{(n)}$ [6], and they are part of a wider statistical methodology called 'Nonparametric Predictive Inference' (NPI) [5, 7]. In the NPI approach, uncertainty is quantified by lower and upper probabilities, which can be regarded as optimal bounds for probabilities based on relatively few assumptions. NPI is a frequentist statistical approach which has strong consistency properties [5] and compares favourably to so-called objective Bayesian methods [7]. Several applications of NPI to problems in statistics, reliability and operations research have been presented, for some references see [1, 7].

Coolen-Schrijner *et al.* [1] considered NPI for system reliability, in particular for series systems with subsystem i a k_i-out-of-m_i system. Such systems are common in practice, and can offer the important advantage of building in redundancy by increasing some m_i to increase the system reliability. Coolen-Schrijner *et al.* [1] applied NPI for Bernoulli data to such systems, with inferences on each subsystem i based on information from tests on n_i components, and the components tested assumed to be exchangeable with the corresponding components to be used in that subsystem. Only situations where components and the system either function or not when called upon were considered. They presented an attractive algorithm for optimal redundancy allocation, with additional components added to subsystems one at a time, which in their setting was proven to be optimal. Hence, NPI for system reliability provides a tractable model, which greatly simplifies optimisation problems involved with redundancy allocation. However, they only proved this result for tests in which no components failed. MacPhee *et al.* [2] generalized this result to redundancy allocation following tests in which any number of components can have failed, a situation in which redundancy plays possibly an even more important role than when testing revealed no failures at all.

In Section 2, an overview of these recent results is given, including a brief introduction to NPI for Bernoulli random quantities. These results are generalized in Section 3 by allowing different k_i-out-of-m_i subsystems to consist of components of the same type, which is an important step towards

NPI for reliability of general systems. Examples in Sections 2 and 3 illustrate the NPI lower and upper probabilities for system functioning. Section 4 concludes the paper with some discussion on the practical relevance of this new theory and corresponding research challenges.

2 NPI FOR SYSTEM RELIABILITY

In this section, NPI for Bernoulli random quantities [3] is summarized, together with the key results for NPI for system reliability by Coolen-Schrijner *et al.* [1] and MacPhee *et al.* [2].

2.1 NPI for Bernoulli quantities

Suppose that there is a sequence of $n+m$ exchangeable Bernoulli trials, each with 'success' and 'failure' as possible outcomes, and data consisting of s successes in n trials. Let Y_1^n denote the random number of successes in trials 1 to n, then a sufficient representation of the data for the inferences considered is $Y_1^n = s$, due to the assumed exchangeability of all trials. Let Y_{n+1}^{n+m} denote the random number of successes in trials $n+1$ to $n+m$. Let $R_t = \{r_1, \ldots, r_t\}$, with $1 \leq t \leq m+1$ and $0 \leq r_1 < r_2 < \ldots < r_t \leq m$, and, for ease of notation, define $\binom{s+r_0}{s} = 0$. Then the NPI upper probability for the event $Y_{n+1}^{n+m} \in R_t$, given data $Y_1^n = s$, for $s \in \{0, \ldots, n\}$, is

$$\overline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = \binom{n+m}{n}^{-1} \times \cdots$$
$$\sum_{j=1}^{t} \left[\binom{s+r_j}{s} - \binom{s+r_{j-1}}{s} \right] \binom{n-s+m-r_j}{n-s}.$$

The corresponding NPI lower probability is derived via the conjugacy property

$$\underline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \in R_t^c | Y_1^n = s)$$

where $R_t^c = \{0, 1, \ldots, m\} \backslash R_t$.

Coolen [3] derived these NPI lower and upper probabilities through direct counting arguments. The method uses the appropriate $A_{(n)}$ assumptions [6] for inference on m future random quantities given n observations, and a latent variable representation with Bernoulli quantities represented by observations on the real line, with a threshold such that successes are to one side and failures to the other side of the threshold. Under these assumptions, the $\binom{n+m}{n}$ different orderings of these observations, when not distinguishing between the n observed values nor between the m future observations, are all equally likely. For each such an ordering, the success-failure threshold can be in any of the $n+m+1$ intervals of the partition of the real line created by the $n+m$ values of the latent variables, leading to $n+m+1$ possible combinations (s, r), with s successes in the n tests and r successes in the m future observations. For such an ordering, these possible (s, r) can be

represented as a path on the rectangular lattice from $(0,0)$ to (n,m) with steps going either one to the right or one upwards. The $\binom{n+m}{n}$ different orderings, which are all equally likely, correspond to the $\binom{n+m}{n}$ different right-upwards paths from $(0,0)$ to (n,m), and hence the above NPI lower and upper probabilities can also be derived by counting paths. To derive the NPI lower probability $\underline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s)$, one counts all such paths which must go through points (s,r) with $r \in R_t$, so they do not go through (s,l) for any $l \in R_t^c$. The corresponding NPI upper probability $\overline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s)$ is derived by counting all such paths that go through at least one (s,r) with $r \in R_t$.

2.2 NPI for a k-out-of-m system

When considering a k-out-of-m system, the event $Y_{n+1}^{n+m} \geq k$ is of interest as this corresponds to successful functioning of such a system, following n tests of components that are exchangeable with the m components in the system. Given data consisting of s successes from n components tested, the NPI lower and upper probabilities for the event that the k-out-of-m system functions successfully are denoted by $\underline{P}(m : k | \, n, s)$ and $\overline{P}(m : k | \, n, s)$, respectively, and these follow from the NPI lower and upper probabilities for $Y_{n+1}^{n+m} \in R_t$ given above. For $k \in \{1, 2, \ldots, m\}$ and $0 < s < n$,

$$\overline{P}(m : k | \, n, s) = \overline{P}(Y_{n+1}^{n+m} \geq k | Y_1^n = s) = \binom{n+m}{n}^{-1} \times \cdots$$
$$\left[\binom{s+k}{s} \binom{n-s+m-k}{n-s} + \sum_{l=k+1}^{m} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right]$$

and, via the conjugacy property,

$$\underline{P}(m : k | \, n, s) = \underline{P}(Y_{n+1}^{n+m} \geq k | Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \leq k-1 | Y_1^n = s)$$
$$= 1 - \binom{n+m}{n}^{-1} \left[\sum_{l=0}^{k-1} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right].$$

For $m = 1$, so considering a system consisting of just a single component, the NPI upper and lower probabilities for the event that the system functions successfully are

$$\overline{P}(1 : 1 | \, n, s) = \overline{P}(Y_{n+1}^{n+1} = 1 | Y_1^n = s) = \frac{s+1}{n+1},$$
$$\underline{P}(1 : 1 | \, n, s) = \underline{P}(Y_{n+1}^{n+1} = 1 | Y_1^n = s) = \frac{s}{n+1}.$$

If the observed data are all successes, so $s = n$, or all failures, so $s = 0$, then

the NPI upper probabilities are, for all $k \in \{1, \ldots, m\}$,

$$\overline{P}(m : k|\, n, n) = \overline{P}(Y_{n+1}^{n+m} \geq k | Y_1^n = n) = 1,$$
$$\overline{P}(m : k|\, n, 0) = \overline{P}(Y_{n+1}^{n+m} \geq k | Y_1^n = 0) = \binom{n+m-k}{n} \binom{n+m}{n}^{-1},$$

and the NPI lower probabilities are, for all $k \in \{1, \ldots, m\}$,

$$\underline{P}(m : k|\, n, n) = \underline{P}(Y_{n+1}^{n+m} \geq k | Y_1^n = n) = 1 - \binom{n+k-1}{n} \binom{n+m}{n}^{-1},$$
$$\underline{P}(m : k|\, n, 0) = \underline{P}(Y_{n+1}^{n+m} \geq k | Y_1^n = 0) = 0.$$

EXAMPLE 1. Table 1 presents the NPI lower and upper probabilities for a k-out-of-62 system, with k varying from 58 to 62, on the basis of tests of n components that are exchangeable with the 62 components in the system, and s components in the tests functioning successfully. If tests have revealed no failures, so $s = n$, then the NPI upper probability of system functioning is equal to 1, which reflects that such tests do not contain evidence against the possibility that such components would always function. The corresponding lower probabilities in these cases are increasing in the number of tests, if the tests did not reveal any failures, which reflects the increasing evidence in favour of at least k components out of 62 functioning in the system. With relatively few tests performed, and many of the 62 components in the system required to function, the effect of a failure in the tests on the predicted system reliability is substantial. This example illustrates that $\underline{P}(m : k|\, n, s) = \overline{P}(m : k|\, n, s-1)$, which generally holds for these NPI lower and upper probabilities [1]. It is worth noticing the lower probability $\underline{P}(62 : 62|\, 62, 62) = 0.5$, which is actually precisely $1/2$ and is the same as would be derived if the whole 62-out-of-62 system were instead considered to be a single unit, and if one exchangeable unit (hence also such a system) had been tested and had been successful, as $\underline{P}(1 : 1|\, 1, 1) = 0.5$.

We return to this example in Section 3 (Example 2), when instead of a single k-out-of-62 system, we regard the system as consisting of two or three k_i-out-of-m_i systems, with the m_i's summing up to 62. Although this example is purely illustrative for the presented theory, the numbers chosen are inspired by the Dutch Oosterscheldekering (Eastern-Scheldt storm surge barrier), which is part of the Delta Works series of dams to protect the Netherlands from flooding. This barrier consists of 62 steel doors, hence the NPI lower and upper probabilities for successful functioning of the system in this example could be interpreted as those for successful functioning of this barrier on a single application, following test results of n doors. Of course, this assumes exchangeability of the functioning of the individual doors, which may not be deemed to be an appropriate assumption.

		$k = 58$		$k = 59$		$k = 60$		$k = 61$		$k = 62$	
n	s	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
1	1	0.079	1	0.063	1	0.048	1	0.032	1	0.016	1
2	2	0.151	1	0.122	1	0.092	1	0.062	1	0.031	1
3	3	0.217	1	0.176	1	0.134	1	0.091	1	0.046	1
	2	0.021	0.217	0.014	0.176	0.008	0.134	0.004	0.091	0.001	0.046
5	5	0.330	1	0.272	1	0.211	1	0.145	1	0.075	1
	4	0.060	0.330	0.041	0.272	0.025	0.211	0.013	0.145	0.005	0.075
10	10	0.538	1	0.458	1	0.367	1	0.260	1	0.139	1
	9	0.192	0.538	0.139	0.458	0.090	0.367	0.049	0.260	0.018	0.139
	8	0.051	0.192	0.032	0.139	0.017	0.090	0.007	0.049	0.002	0.018
	7	0.011	0.051	0.006	0.032	0.003	0.017	0.001	0.007	0.000	0.002
20	20	0.763	1	0.681	1	0.573	1	0.431	1	0.244	1
30	30	0.868	1	0.800	1	0.699	1	0.548	1	0.326	1
50	50	0.952	1	0.910	1	0.834	1	0.696	1	0.446	1
60	60	0.969	1	0.936	1	0.872	1	0.744	1	0.492	1
62	62	0.971	1	0.941	1	0.878	1	0.752	1	0.500	1
100	100	0.993	1	0.980	1	0.946	1	0.855	1	0.617	1

TABLE 1. NPI lower and upper probabilities for a k-out-of-62 system

2.3 k_i-out-of-m_i subsystems in series configuration

Many systems consist of series configurations of $N \geq 2$ independent subsystems, with subsystem i ($i = 1, \ldots, N$) a k_i-out-of-m_i system consisting of exchangeable components. Assume that, in relation to subsystem i, n_i components that are exchangeable with those to be used in the subsystem have been tested, of which s_i functioned successfully. For the series system to function, all its subsystems must function, and due to the assumed independence of the subsystems (which implies independence of components in different subsystems), the NPI upper and lower probabilities for such a series system to function are

$$\overline{P}(\boldsymbol{m} : \boldsymbol{k}|\, \boldsymbol{n}, \boldsymbol{s}) = \textstyle\prod_{i=1}^{N} \overline{P}(m_i : k_i|\, n_i, s_i),$$
$$\underline{P}(\boldsymbol{m} : \boldsymbol{k}|\, \boldsymbol{n}, \boldsymbol{s}) = \textstyle\prod_{i=1}^{N} \underline{P}(m_i : k_i|\, n_i, s_i),$$

where the notation with N-vectors $\boldsymbol{m}$, $\boldsymbol{k}$, $\boldsymbol{n}$, $\boldsymbol{s}$ has been introduced to generalize earlier notation. Coolen-Schrijner *et al.* [1] presented a powerful algorithm for optimal redundancy allocation for such systems, that is how best to assign additional components to subsystems (hence to increase the number of components m_i), for situations where the required numbers of components that must function for the subsystems remains the same (k_i). They only considered such redundancy allocation after zero-failure testing (so $s_i = n_i$ for all $i = 1, \ldots, N$). MacPhee *et al.* [2] succeeded in generalizing this algorithm to general test results. In these papers, the NPI lower probability for system functioning was used as the reliability measure. We do not discuss the redundancy allocation algorithm in this paper, but will present the NPI lower and upper probabilities for functioning of a system consisting of multiple k_i-out-of-m_i subsystems in a series configuration, with all subsystems consisting of the same type of component. This is an important step towards developing NPI for reliability of general systems.

3 MULTIPLE SUBSYSTEMS WITH ONE TYPE OF COMPONENT

The results summarized in Section 2 need to be generalized in order to develop the NPI framework for reliability of more general systems. As a first important step, we consider how to deal with exchangeable components appearing in different subsystems. Such components are exchangeable as far as learning from test results is concerned, but they have different roles in the overall system hence they must be distinguished. In the NPI approach, the interdependence of the components to be used in the system is explicitly taken into account, and we need to generalize the results by Coolen [3] for the situation where the m future components belong to different subgroups, with required numbers of successes specified per subgroup.

We present this generalization here for a series system consisting of two subsystems, with subsystem $i = 1, 2$ a k_i-out-of-m_i system, and both these subsystems consisting of the same type of component. As before, we assume that n components which are exchangeable with the m_1 and m_2 components in these subsystems have been tested, and that s of these functioned successfully. This system will function if at least k_1 of the m_1 components in subsystem 1 function, together with at least k_2 of the m_2 components in subsystem 2. The NPI lower probability for this event is

$$\underline{P}(m_1 : k_1, m_2 : k_2 \mid n, s) = \binom{n + m_1 + m_2}{n, m_1, m_2}^{-1} \times \cdots \sum_{l_1=k_1}^{m_1} \sum_{l_2=k_2}^{m_2} \binom{s - 1 + l_1 + l_2}{s - 1, l_1, l_2} \binom{n - s + m_1 - l_1 + m_2 - l_2}{n - s, m_1 - l_1, m_2 - l_2}$$

and the corresponding NPI upper probability is

$$\begin{aligned}\overline{P}(m_1 : k_1, m_2 : k_2 \mid n, s) = {} & \binom{n + m_1 + m_2}{n, m_1, m_2}^{-1} \times \cdots \\ & \left[\sum_{l_1=k_1}^{m_1} \binom{s + l_1 + k_2 - 1}{s, l_1, k_2 - 1} \binom{n - s + m_1 - l_1 + m_2 - k_2}{n - s, m_1 - l_1, m_2 - k_2} + \cdots \right. \\ & \sum_{l_2=k_2}^{m_2} \binom{s + k_1 - 1 + l_2}{s, k_1 - 1, l_2} \binom{n - s + m_1 - k_1 + m_2 - l_2}{n - s, m_1 - k_1, m_2 - l_2} + \cdots \\ & \left. \sum_{l_1=k_1}^{m_1} \sum_{l_2=k_2}^{m_2} \binom{s - 1 + l_1 + l_2}{s - 1, l_1, l_2} \binom{n - s + m_1 - l_1 + m_2 - l_2}{n - s, m_1 - l_1, m_2 - l_2} \right].\end{aligned}$$

These NPI lower and upper probabilities are derived by counting paths on the grid from $(0, 0, 0)$ to (n, m_1, m_2), in a similar way as described in Section 2. By the appropriate $A_{(n)}$ assumptions, all orderings of the $n + m_1 + m_2$ latent variables representing the n test observations and the m_1 and

m_2 future random quantities are again equally likely, and each such ordering can again be represented by a unique path from $(0,0,0)$ to (n, m_1, m_2). The above NPI lower probability follows by counting all paths which go through (s, r_1, r_2) for $r_1 \geq k_1$ and $r_2 \geq k_2$ but not through any point (s, r_1, r_2) with r_1 less than k_1 or with r_2 less than k_2, The corresponding NPI upper probability follows by counting all such paths that go through at least one point (s, r_1, r_2) with $r_1 \geq k_1$ and $r_2 \geq k_2$.

These results have been generalized to systems with $L > 2$ k_i-out-of-m_i subsystems in a series configuration, by using similar counting arguments on an $L+1$-dimensional grid. Due to space limitations, the general results will be presented elsewhere, together with more detailed justification of the arguments underlying these NPI lower and upper probabilities. However, in Example 2 we briefly illustrate a case related to that presented in Example 1 in Section 2, but with the system split up into two or three subsystems. For this, we will use the following NPI lower and upper probabilities for system functioning for the case with $L = 3$:

$$\underline{P}(m_1 : k_1, m_2 : k_2, m_3 : k_3 \mid n, s) = \binom{n+m_1+m_2+m_3}{n, m_1, m_2, m_3}^{-1} \times \cdots$$

$$\sum_{l_1=k_1}^{m_1} \sum_{l_2=k_2}^{m_2} \sum_{l_3=k_3}^{m_3} \binom{s-1+l_1+l_2+l_3}{s-1, l_1, l_2, l_3} \binom{n-s+m_1-l_1+m_2-l_2+m_3-l_3}{n-s, m_1-l_1, m_2-l_2, m_3-l_3},$$

$$\overline{P}(m_1 : k_1, m_2 : k_2, m_3 : k_3 \mid n, s) = \binom{n+m_1+m_2+m_3}{n, m_1, m_2, m_3}^{-1} \times \cdots$$

$$\left[\sum_{l_2=k_2}^{m_2} \sum_{l_3=k_3}^{m_3} \binom{s+k_1-1+l_2+l_3}{s, k_1-1, l_2, l_3} \binom{n-s+m_1-k_1+m_2-l_2+m_3-l_3}{n-s, m_1-k_1, m_2-l_2, m_3-l_3} + \right.$$

$$\sum_{l_1=k_1}^{m_1} \sum_{l_3=k_3}^{m_3} \binom{s+l_1+k_2-1+l_3}{s, l_1, k_2-1, l_3} \binom{n-s+m_1-l_1+m_2-k_2+m_3-l_3}{n-s, m_1-l_1, m_2-k_2, m_3-l_3} +$$

$$\sum_{l_1=k_1}^{m_1} \sum_{l_2=k_2}^{m_2} \binom{s+l_1+l_2+k_3-1}{s, l_1, l_2, k_3-1} \binom{n-s+m_1-l_1+m_2-l_2+m_3-k_3}{n-s, m_1-l_1, m_2-l_2, m_3-k_3} +$$

$$\left. \sum_{l_1=k_1}^{m_1} \sum_{l_2=k_2}^{m_2} \sum_{l_3=k_3}^{m_3} \binom{s-1+l_1+l_2+l_3}{s-1, l_1, l_2, l_3} \binom{n-s+m_1-l_1+m_2-l_2+m_3-l_3}{n-s, m_1-l_1, m_2-l_2, m_3-l_3} \right].$$

If testing revealed no failing components, so $s = n$, then these NPI upper probabilities, for any number L of subsystems, are equal to 1 for all values of m_i and k_i, which again reflects that such test data do not provide strong evidence against the possibility that such components would never fail.

Example 2. We return to the situation described in Example 1, inspired by the number of steel doors in the Oosterscheldekering. Actually, instead of one line of 62 doors next to each other, the barrier consists of three sections, with 15 steel doors in the northern section, 16 in the middle section, and 31 in the southern section. Suppose now that the functioning of

the barrier requires specific numbers of doors in each section to function. While the assumption of exchangeability of the doors remains with regard to the uncertainty of their functioning and the way in which we learn from test data on similar doors, for the functioning of the system it is important to distinguish the doors according to which section they are in. For this, the theory in this section is suitable. First, let us suppose that the northern and middle sections can be combined to one k_1-out-of-31 subsystem, with the southern section a separate k_2-out-of-31 subsystem, and these two subsystems form together the overall system in series configuration. Some NPI lower and upper probabilities for functioning of the whole system are presented in Table 2.

		$k_1 = k_2 = 29$		$k_1 = 29, k_2 = 30$		$k_1 = k_2 = 30$		$k_1 = k_2 = 31$	
n	s	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
1	1	0.066	1	0.050	1	0.040	1	0.016	1
2	2	0.126	1	0.096	1	0.077	1	0.031	1
3	3	0.182	1	0.139	1	0.113	1	0.046	1
	2	0.015	0.182	0.010	0.139	0.006	0.113	0.001	0.046
5	5	0.280	1	0.218	1	0.178	1	0.075	1
	4	0.045	0.280	0.028	0.218	0.019	0.178	0.005	0.075
10	10	0.467	1	0.375	1	0.314	1	0.139	1
	9	0.148	0.467	0.099	0.375	0.070	0.314	0.018	0.139
	8	0.036	0.148	0.020	0.099	0.012	0.070	0.002	0.018
	7	0.007	0.036	0.003	0.020	0.002	0.012	0.000	0.002
20	20	0.687	1	0.579	1	0.503	1	0.244	1
30	30	0.803	1	0.701	1	0.625	1	0.326	1
50	50	0.908	1	0.829	1	0.766	1	0.446	1
60	60	0.934	1	0.865	1	0.809	1	0.492	1
62	62	0.938	1	0.871	1	0.816	1	0.500	1
100	100	0.977	1	0.936	1	0.901	1	0.617	1

TABLE 2. NPI lower and upper probabilities with $m_1 = m_2 = 31$

Comparing Tables 1 and 2, it is clear that the lower and upper probabilities in the final columns, where the system only functions if all components function, are identical. This is logical, as in both cases it just means that, after n components have been tested, the next m components must all function. The three other cases presented in Table 2 do not directly relate to cases in Table 1, due to the different system configurations. Clearly, a 60-out-of-62 system can function for more combinations of failing components than two 30-out-of-31 subsystems in a series configuration, namely the former still functions if the two failing components happen to be in the same subsystem corresponding to it, in which case the latter would not function anymore. This explains why the entries (except those equal to 1) in Table 1 are greater than corresponding ones in Table 2, where we relate the cases $k = 60$ with $k_1 = k_2 = 30$ and also $k = 58$ with $k_1 = k_2 = 29$.

Let us now consider the system of 62 components split up into three subsystems, with $m_1 = 15$, $m_2 = 16$ and $m_3 = 31$ components, inspired by the three sections of the Oosterscheldekering. First, let us consider the reliability of each of these three subsystems independently of each other, so we consider each as a single k-out-of-m system. The NPI lower and upper

probabilities for successful functioning of each of these systems individually, based on s successfully functioning components in n tests, are given in Table 3, for the values k and m as indicated in the columns.

		$k=15, m=15$		$k=16, m=16$		$k=30, m=31$		$k=31, m=31$	
n	s	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
1	1	0.063	1	0.059	1	0.063	1	0.031	1
2	2	0.118	1	0.111	1	0.119	1	0.061	1
3	3	0.167	1	0.158	1	0.171	1	0.088	1
	2	0.020	0.167	0.018	0.158	0.016	0.171	0.005	0.088
5	5	0.250	1	0.238	1	0.262	1	0.139	1
	4	0.053	0.250	0.048	0.238	0.045	0.262	0.016	0.139
10	10	0.400	1	0.385	1	0.433	1	0.244	1
	9	0.150	0.400	0.138	0.385	0.142	0.433	0.055	0.244
	8	0.052	0.15	0.046	0.138	0.039	0.142	0.011	0.055
	7	0.017	0.052	0.014	0.046	0.009	0.039	0.002	0.011
20	20	0.571	1	0.556	1	0.635	1	0.392	1
30	30	0.667	1	0.651	1	0.746	1	0.492	1
50	50	0.769	1	0.758	1	0.856	1	0.617	1
60	60	0.800	1	0.789	1	0.886	1	0.659	1
62	62	0.805	1	0.795	1	0.891	1	0.667	1
100	100	0.870	1	0.862	1	0.945	1	0.763	1

TABLE 3. NPI lower and upper probabilities for k-out-of-m systems

These NPI lower and upper probabilities give an indication of the reliability of the individual subsystems considered, when considering them independently of the other systems. It is crucial, however, that in the application considered in this example, these subsystems consist of the same type of component, for which only limited test information is available. Hence, if it were known that one of these subsystems functions satisfactorily, let us assume this would be the subsystem with $m = 15$ and assuming that this would function only if $k = 15$, then for the next subsystem considered we are more confident in the reliability of the components, as now in addition to the test results for the n tested components it is known that a further 15 components all function satisfactorily. This has a substantial impact on overall reliability when we combine the subsystems into a single system.

If one were to neglect the interdependence of the components in the different subsystems, one would make the mistake of quantifying the system's reliability by multiplying the NPI lower and upper probabilities of successful functioning of the subsystems, as briefly mentioned in Section 2 for independent subsystems. For example, consider the third column of Table 2, involving a series system with two 30-out-of-31 subsystems on the basis of s components functioning well out of n components tested. If we would, instead, multiply the lower and upper probabilities for two individual 30-out-of-31 systems, based on the same test information, so effectively we would take the squared values of the entries in the third column in Table 3, then the latter would lead to substantially smaller values for the lower probability, and also for the upper probability for all cases where this is not equal to one. To illustrate this important issue, assume that $n = 10$ components had been tested, of which $s = 9$ functioned successfully. The

corresponding NPI lower and upper probabilities for successful functioning of the series system with two 30-out-of-31 subsystems (Table 2, third column) are 0.070 and 0.314, respectively. If one would, mistakenly, neglect the interdependence of these two subsystems, which use components of the same type, and multiply the NPI lower and upper probabilities for the individual 30-out-of-31 subsystems (Table 3, third column), this would lead to the values $0.142^2 = 0.020$ for the lower and $0.433^2 = 0.187$ for the upper probability, which are substantially smaller than the correct values.

Let us now consider the 62-component system as consisting of three subsystems in series structure, with $m_1 = 15$, $m_2 = 16$ and $m_3 = 31$ components. Table 4 presents NPI lower and upper probabilities for some situations reflecting satisfactory functioning of the whole system depending on the specific numbers k_i $(i = 1, 2, 3)$ of components required to function per subsystem.

(k_1, k_2, k_3) :		$(14, 15, 30)$		$(15, 16, 30)$		$(15, 16, 31)$	
n	s	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
1	1	0.045	1	0.024	1	0.016	1
2	2	0.087	1	0.047	1	0.031	1
3	3	0.127	1	0.069	1	0.046	1
	2	0.008	0.127	0.003	0.069	0.001	0.046
5	5	0.197	1	0.110	1	0.075	1
	4	0.024	0.197	0.009	0.110	0.005	0.075
10	10	0.345	1	0.200	1	0.139	1
	9	0.085	0.345	0.033	0.200	0.018	0.139
	8	0.016	0.085	0.005	0.033	0.002	0.018
	7	0.003	0.016	0.001	0.005	0.000	0.002
20	20	0.542	1	0.337	1	0.244	1
30	30	0.664	1	0.437	1	0.326	1
50	50	0.799	1	0.571	1	0.446	1
60	60	0.838	1	0.618	1	0.492	1
62	62	0.844	1	0.626	1	0.500	1
100	100	0.919	1	0.736	1	0.617	1

TABLE 4. NPI lower and upper probabilities with $m_1 = 15, m_2 = 16, m_3 = 31$

Again, if all 62 components need to function ($k_i = m_i$ for all i), then the NPI lower and upper probabilities are as in Tables 1 and 2 for the same situation. Suppose that the whole system functions satisfactorily if in each subsystem not more than one component fails, leading to the NPI lower and upper probabilities in the first column of Table 4. If we had not separated the two smallest subsystems, so instead had assumed that the whole system consisted of two subsystems with $m_1 = m_2 = 31$, as considered at the start of this example with corresponding NPI lower and upper probabilities given in Table 2, and if we had allowed two failing components for the first subsystem with $m_1 = 31$ components, then (see column 2 in Table 2) the NPI lower and upper probabilities (the latter if different from 1) would have been larger than those with the three subsystems taken into account separately. This is due to the fact that there would be more combinations of the failing components included in the counts for the lower and upper probabilities in Table 2, namely those with two failing components in one, and zero in

the other, of the individual subsystems with 15 and 16 components. This illustrates clearly that one must carefully define the requirements on the subsystems in order for the overall system to function, which is of course directly linked to the appropriate system structure.

Examples 1 and 2 clearly show the effect of increasing numbers of tests on the system reliability. If all n components tested succeeded in their task, so $s = n$, then the NPI lower probabilities increase as function of n, but the rate of increase decreases. This is in line with intuition as it reflects that, with all tests being successful, the positive effect of a further successful test on the lower probability of system functioning decreases with increasing n. This can also be used to set a minimum number of tests, assuming no failures will be discovered, in order to meet a reliability requirement formulated as a minimum value for the NPI lower probability of system functioning. This is relevant in high-reliability testing, where failures in tests typically lead to redesign of the units followed by a new stage of testing, and hence one needs to determine how many zero-failure tests are required in order to demonstrate reliability. Coolen and Coolen-Schrijner [8, 9] present related theory and methods from the perspectives of NPI and Bayesian statistics.

4 DISCUSSION

The NPI approach to system reliability is in early stages of development. It provides a new method for statistical inference on system reliability on the basis of limited information resulting from component testing. In the reliability literature, system reliability is usually expressed as function of failure probabilities for components, which are typically assumed to be known. Under limited information, this will clearly not be the case, and the proper inclusion of uncertainty about components' failure probabilities is rarely addressed. One cannot replace parameters representing such failure probabilities by estimates, as the system reliability function is typically non-linear. More importantly, any such classical approach with parameters representing components' failure probabilities does not take into account the interdependence of the components to be used in the system of interest.

One can use a Bayesian approach, expressing the system reliability via a posterior predictive distribution, which will take care of this interdependence, but this requires the use of prior distributions for the parameters, which adds further assumptions that may be hard to justify. This is particularly clear when considering system reliability after zero-failure tests, where Bayesian methods will typically lead to a probability of system functioning that is less than one, while clearly the test data do not strongly suggest that components might actually fail. The use of lower and upper probabilities in reliability is attractive in such situations as the upper probability of system functioning, given no test failures, can be equal to one (as the NPI upper probabilities are), reflecting no evidence that things can go wrong. In such cases, the corresponding lower probability may be of most use, as it re-

flects the amount of evidence available in favour of system functioning, and as it enables cautious inference which is often deemed appropriate in risk analysis. The fact that the NPI lower and upper probabilities result from combinatorial arguments, based only on an exchangeability assumption and an underlying latent variable representation is also attractive.

This paper presents an important step in the development of NPI for more complex system structures, as components of one type frequently occur in different subsystems. The next challenge is development of NPI for k-out-of-m systems which contain different types of components, which is not straightforward due to the use of lower and upper probabilities. Although the development of NPI for system reliability is still far from the point where it can be applied to substantial practical systems, the results for small systems clearly show the importance of such an approach which implicitly takes limited information on component reliability into account.

In two recent papers [1, 2] a powerful optimal algorithm was presented for redundancy allocation related to the NPI approach to reliability of systems consisting of independent k_i-out-of-m_i subsystems, each consisting of a single type of component, which are different for different subsystems. A myopic algorithm was proven to be optimal, and this algorithm is straightforward to implement and requires negligible computing time. Research is currently ongoing to justify a similar algorithm for the scenario discussed in this paper. Numerical examples indicate that a similarly attractive algorithm will again be optimal, but proving this property is rather complicated.

NPI lower and upper probabilities for system reliability are based on combinatorics, so the computation time will increase for more substantial systems. However, there are no complex integrals involved (as e.g. is typically the case in Bayesian statistics), and as all sums are finite there are no major difficulties. For large systems it may be required to consider approximations for the sums involved in deriving the NPI lower and upper probabilities, but NPI is not yet developed to the stage where this has become relevant. If more test data become available, updating the NPI lower and upper probabilities occurs by calculating them again using all combined information, there is no straightforward sequential updating algorithm available as is the case in Bayesian statistics. In fact, updating in NPI is explicitly not the same as conditioning, see Augustin and Coolen [5] for more detailed discussion of this important foundational aspect, together with consistency results under updating for NPI for real-valued random quantities. The imprecision, that is the difference between corresponding NPI upper and lower probabilities, tends to decrease as a function of n and increase as a function of m, although the imprecision tends to become smaller for non-trivial events if both the upper and lower probabilities get close to either zero or to one. It will be of interest to study this in more detail, in particular as imprecision seems to relate logically to the amount of information available and to the number of future random quantities involved in the event of interest.

Although a nonparametric approach as presented in this paper is attractive, it has obvious limitations. For example, if NPI were developed further in order to take ageing of technical components into account, the huge amount of data needed to describe the effects of ageing without the use of a parametric model will make the approach of little practical value. One of the main research challenges for NPI will be to combine it with partial parametric modelling to model aspects of ageing using specific processes [10]. This may lead to a novel semi-parametric approach that could be of benefit to a wide range of applications. The use of lower and upper probabilities in combination with stochastic processes is an exciting topic area for future research, which has not attracted much attention so far.

The use of lower and upper probabilities is attractive for many problems in reliability, as they can deal more explicitly with limited information. Utkin and Coolen [4] present an introductory overview of many methods and applications presented, mostly during the past decade. This also includes references to other applications of NPI in reliability.

Bibliography

[1] P. Coolen-Schrijner, F. P. A. Coolen, and I. M. MacPhee. Nonparametric predictive inference for system reliability with redundancy allocation. *Journal of Risk and Reliability*, 222:463–476, 2008.

[2] I. M. MacPhee, F. P. A. Coolen, and A. M. Aboalkhair. Nonparametric predictive system reliability with redundancy allocation following component testing. *Journal of Risk and Reliability*, 223:181–188, 2009.

[3] F. P. A. Coolen. Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36:349–357, 1998.

[4] L. V. Utkin and F. P. A. Coolen. Imprecise reliability: an introductory overview. In G. Levitin, editor, *Computational Intelligence in Reliability Engineering, Volume 2: New Metaheuristics, Neural and Fuzzy Techniques in Reliability*, pages 261–306. Springer, 2007.

[5] T. Augustin and F. P. A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124:251–272, 2004.

[6] B. M. Hill. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63:677–691, 1968.

[7] F. P. A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15:21–47, 2006.

[8] F. P. A. Coolen and P. Coolen-Schrijner. Nonparametric predictive reliability demonstration for failure-free periods. *IMA Journal of Management Mathematics*, 16:1–11, 2005.

[9] F. P. A. Coolen and P. Coolen-Schrijner. Bayesian reliability demonstration. In F. Ruggeri, R. Kenett, and F. W. Faltin, editors, *Wiley Encyclopedia of Statistics in Quality and Reliability*, pages 196–202. John Wiley & Sons, Chichester, 2007.

[10] J. M. van Noortwijk. A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1):2–21, 2009.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Multi-criteria optimization of life-cycle performance of structural systems under uncertainty

Dan M. Frangopol* and Nader M. Okasha
– Lehigh University, Bethlehem, Philadelphia, USA

Abstract. Prediction of the life-cycle performance of structural systems must be accompanied with an efficient intervention planning procedure that assures the safe upkeep of structures. Multi-criteria optimization is an effective approach for conducting this procedure. Life-cycle performance of structural systems is typically quantified by means of performance indicators. The ability of the performance measures and their predictive models to accurately interpret and quantify the effects of applying maintenance interventions is necessary. The objective of this paper is to review recent advances in methods of multi-criteria optimization of life-cycle performance of structural systems under uncertainty. Two approaches for finding optimum maintenance strategies for deteriorating structural systems through multi-criteria optimization and using genetic algorithms are presented with applications. These approaches use different problem formulations and types of performance indicators.

1 INTRODUCTION

In their paper, the use of lifetime distributions in bridge maintenance and replacement modelling, van Noortwijk and Klatter [1] recognized the importance of life-cycle analysis for the optimization of management of roads and bridges. They also acknowledged that "to calculate the life-cycle cost, information on the time and cost of bridge maintenance and replacement is needed". It is of evident necessity that proper modeling procedures are implemented for the accurate prediction of the times of bridge maintenance and replacement under uncertainty. Knowledge of these times establishes the basis for optimizing the proper and most economical maintenance procedures required. These times are typically assesed by means of performance indicators. Accordingly, a maintenance optimization problem is formulated. Tools for solving the optimization problem efficiently are needed. In the past decade, the maintenance optimization problem has usually been formulated

*corresponding author: Department of Civil and Environmental Engineering, Center for Advanced Technology for Large Structural Systems (ATLSS Center) Lehigh University, 117 ATLSS Drive, Imbt Labs, Bethlehem, PA 18015-4729, U.S.A. telephone: +1-(610) 758 6103, fax: +1-(610) 758 4115, e-mail: dan.frangopol@lehigh.edu

as a multi-criteria one, and the tool of choice for solving this optimization problem has become the genetic algorithms.

In this paper, recent advances in methods of multi-criteria optimization of life-cycle performance of structural systems under uncertainty are briefly reviewed. Two approaches for finding optimum maintenance strategies for deteriorating structural systems through multi-criteria optimization and using genetic algorithms are presented with applications. These approaches use different problem formulations and types of performance indicators

2 PERFORMANCE INDICATORS

Upkeep of the safe structural performance is the primary goal of any maintenance procedure. By doing so, the service life of structures may, in fact, be further prolonged. In order to keep track of the structural performance, inidcators that represent different types of the structural performance are developed and used as the main tool in deciding the timing of appliction of maintenance [2]. In his paper, coauthered with Frangopol and Kallen [3], van Noortwijk reviewed different types of models for prediction of structural performance. These models were classified as random-variable models, such as the reliability index, the time-dependent reliability index, and the failure rate; and stochastic process models such as the Markov decision processes and the renewal models [3].

Because of the aleatory and epistemic uncertainties, structural reliability has been a major decision factor throughout the life-cycle of engineering structures. The reliability index was shown to be a good tool for prioritizing the maintenance actions [4]. Enright and Frangopol [5] have utilized the cumulative-time probability of failure in maintenance planning. Other reliability-oriented performance indices have also been implemented in maintenance planning. For instance, Yang *et al.* [6, 7] have used lifetime functions that quantify the survivability and hazard rates of the structures. Other indicators, particularly the safety and condition indices have heavily been used in life-cycle management and maintenance optimization [8, 9, 10, 11]. Recently, Okasha and Frangopol [12] have pointed out the importance of integrating the redundancy of structures as an additional decision tool in the maintenance optimization process.

3 GENETIC ALGORITHMS

Inspired by evolutionary biology, genetic algorithms (GAs) have found their way into a large number of optimization applications and the growing interest in them continues. This is due to several advantages of GAs compared to other methods for complex problems. It is enough to be able to evaluate the objective functions for a given set of input parameters in order to solve a certain optimization problem using GAs. In addition, GAs are especially attractive in solving multi-criteria problems due to their ability of finding a set of Pareto-optimal solutions in one run compared to conventional meth-

ods that can only find one solution per run. In the past decade, GAs have been the method of choice for solving multi-objective maintenance optimization problems. In particular, a GA algorithm called non-dominated sorting GA with controlled elitism, NSGA-II [13] has been the most widely used algorithm for these applications.

This NSGA-II algorithm can be briefly described as follows. An initial (parent) population is randomly generated. Non-dominated sorting is performed in order to provide a measure of fitness and locate the individuals in fronts, where the first front is a potential Pareto-optimal. A set of operations are performed next for a specified number of generations. In each generation, binary tournament selection, cross-over and mutation are performed to generate an offspring population that is combined with the parent population and from which the best individuals are selected to pass through the next generation.

4 MAINTENANCE OPTIMIZATION

In this paper, two distinct approaches for the multi-objective optimization of maintenace are presented. These two approaches differ mainly in their formulation, performance indicators used and target application they are intended for, but are both solved using the NSGA algorithm. The performance indices considered in the first approach are the instantaneous probability of system failure, redundancy, and life-cycle cost (LCC), whereas in the second approach, they are the unavailability, redundancy, and LCC. Both approaches are illustrated with examples.

4.1 Approach 1

The first maintenance optimization approach is applied to a five-bar truss under a horizontal random load. The bars are grouped into the three groups of equal areas A_1, which includes the two vertical bars, A_2 which includes the horizontal bar, and A_3 which includes the two diagonal bars.

A detailed description of the truss, the random variables and load and resistance models can be found in Okasha and Frangopol [14]. The design variables considered are a maintenance code M, (a binary variable with three bits), and the application time variable t. Each bit of M represents one of the three bar groups considered. Each group has its bars replaced if the corresponding bit in M takes a value of 1 and not replaced if the value is 0. The cost of replacing bar groups A_1, A_2 and A_3 are assumed, respectively, as \$1800, \$900, \$2550. Constraints are imposed on the thresholds for the instantanous probability of system failure and redundancy, and time of application of the mainenance actions. Accordingly, the formulation of the problem may be stated as follows:

Find:

- The time of application of maintenance: t, $(1a)$

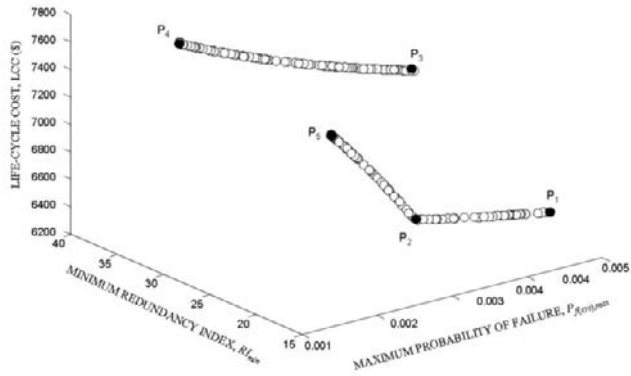


FIGURE 1. Pareto-optimal set for a five-bar truss (adapted from [14])

- The maintenance code: M, $(1b)$

To achieve the following three objectives:

- Minimize $P_{f(sys),max}$, $(1c)$
- Maximize RI_{min}, $(1d)$
- Minimize LCC, $(1e)$

Subject to the constraints:

- $P_{f(sys),max} \leq P_{f(sys),allowable}$, $(1f)$
- $RI_{min} \geq RI_{allowable}$, $(1g)$
- $5 \leq t \leq 45$ years, $(1h)$

where $P_{f(sys),max}$ is the maximum (worst) value reached for the probability of system failure throughout the service life, RI_{min} is the minimum (worst) value reached for the redundancy index throughout the service life, $P_{f(sys),allowable}$ is the allowable maximum probability of system failure, and $RI_{allowable}$ is the allowable minimum redundancy index.

Figure 1 shows the Pareto-optimal set obtained. Projections of the results presented in Figure 1 in the bidimensional space are presented in Figure 2. Five solutions selected from the Pareto-optimal set in Figures 1 and 2 are investigated. The history profiles for reliability, redundancy, LCC and bar areas associated with these five solutions are shown in Figure 3.

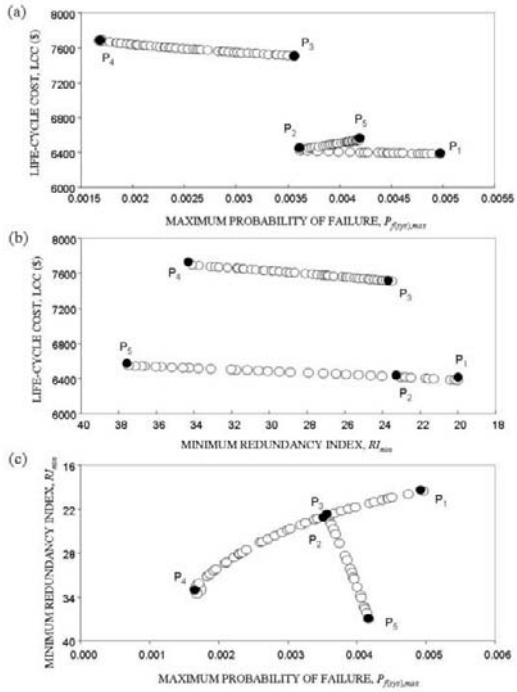


FIGURE 2. Projections of the Pareto-optimal set for a five-bar truss in each of the three bidimensional spaces (adapted from [14])

It is found that in Solution P_1 replacing the horizontal and vertical bars (groups 1 and 2) is enough to maintain the safety and redundancy of the structure at this time, and thus, replacing the diagonal bars (group 3) is not necessary and will only result in unnecessary expenses. The $P_{f(sys),max}$ and RI_{min} values obtained by Solution P_2 are the best that can be achieved with $M = [110]$, i.e. without replacing the diagonal bars. Solution P_3 shows that further improvement in $P_{f(sys),max}$ cannot be achieved without replacing all bars. For this reason a jump exists in the LCC from solutions P_2 to P_3 and the Pareto-optimal curve is discontinuous between these solutions. See Okasha and Frangopol [14] for further details.

4.2 Approach 2

The second maintenance optimization approach is applied to the superstructure of the Colorado Bridge E-17-AH. A detailed description of this bridge can be found in [4]. The bridge has three spans of equal length. The reinforced concrete slab is supported by nine standard-rolled, compact, noncomposite steel girders [4]. As shown in Figure 4, the system failure is

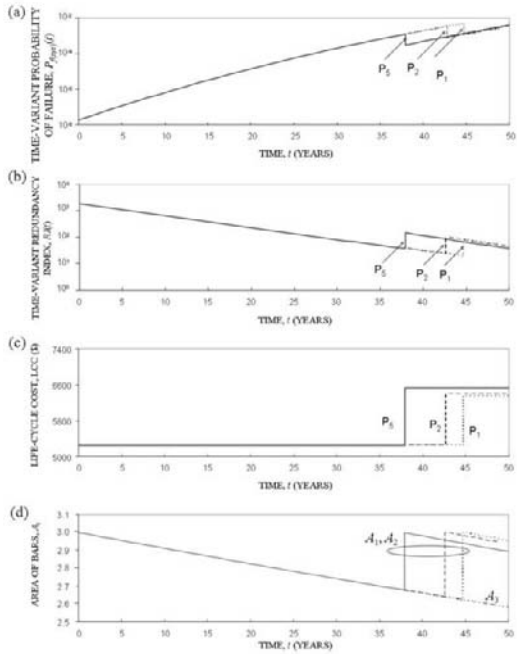


FIGURE 3. History profiles for: (a) area of vertical bars A_1; (b) area of horizontal bars A_2 and; (c) area of diagonal bars A_3 for selected optimum solutions

assumed to occur by the failure of any three adjacent girders or the deck. In Figure 4, the deck is denoted as D and the girders $1, 2, ..., 9$ are denoted as $G1, G2, ..., G9$, respectively.

Four essential maintenance options are considered and the target service life is 75 years. The essential maintenance actions and their associated costs are [4]: Replace deck ($225,600); Replace exterior girders ($229,200); Replace exterior girders and deck ($341,800); and Replace superstructure ($487,100). For preventive maintenance, silane treatment is considered for maintaining the deck and re-painting is considered for maintaining the girders. The cost of silane treatment for the entire deck is assumed as $50,000 and the cost of girder re-painting for all girders is assumed as $100,000 [15].

An essential maintenance is applied at the time a performance threshold is reached, where the type of maintenance applied is the one that provides the lowest present cost per year of increase of service life [4]. Preventive maintenance is applied based on the results of the optimization design variables. The design variables considered are: a continuous design variable for the unavailability threshold An_{th}, a continuous design variable for the redundancy threshold RI_{th}, ten continuous design variables for the time of

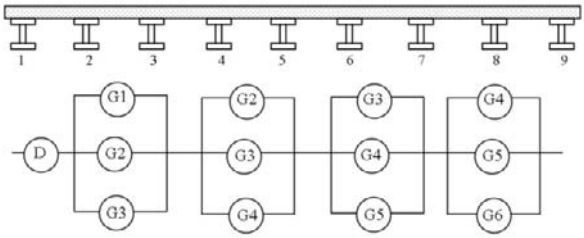


FIGURE 4. Series-parallel models of Bridge E-17-AH

application of the preventive maintenance of the deck Td_i $(i = 1, 2, ..., 10)$, an integer design variable for the number of applications of the preventive maintenance for the deck Nd (where $Nd = 0, 1, 2, ..., 10$), ten continuous design variables for the time of application variables for the preventive maintenance of the girders Tg_i $(j = 1, 2, ..., 10)$; and an integer design variable for the number of applications of the preventive maintenance for the girders Ng (where $Ng = 0, 1, 2, ..., 10$). Constraints are imposed on the thresholds for the unavailability and redundancy, and time of application of the mainenance actions.

Accordingly, the formulation of the problem is stated as follows [15]:

Find: An_{th}, RI_{th}, Td_i, Nd, Tg_i, Ng *to achieve the following three objectives:*

- Minimize An_{max}, $(2a)$
- Maximize RI_{min}, $(2b)$
- Minimize LCC, $(2c)$

Subject to the constraints:

- $10^{-1} \leq An_{th} \leq 10^{-3}$, $(2d)$
- $10^{1} \leq RI_{th} \leq 10^{4}$, $(2e)$
- $2 \leq Tm_i \leq 73$, $(2f)$
- $Tm_i - Tm_{i-1} \geq 2$, $(2g)$
- $Nd = 0, 1, 2, ..., 10$, $(2h)$
- $Ng = 0, 1, 2, ..., 10$, $(2i)$

where An_{max} is the maximum (worst) value reached for the unavailabliity throughout the service life, RI_{min} the minimum (worst) value reached for the redundancy index throughout the service life, and Tm_i is the time of maintenance application i, and $i, j = 1, 2, ..., 10$.

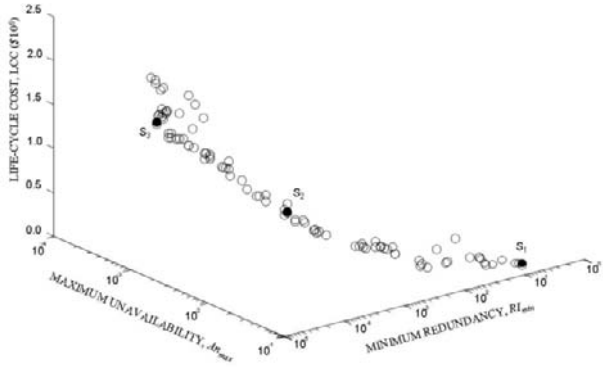


FIGURE 5. Pareto-optimal sets of for Bridge E-17-AH (adapted from [15])

The resulting three dimensional Pareto-optimal set of the optimization problem is shown in Figure 5. Projections of this set in the bidimensional space are presented in Figure 6. Each point in Figures 5 and 6 represents an optimum maintenance plan. A choice among these solutions can be made by decision makers based on their budgets and preferences. This choice will be guided by the trends observed in the figures. For example, as the unavailability is reduced and/or the redundancy is increased, the associated LCC is increased. However, the increase in LCC is relatively higher with reducing the unavailability than with increasing the redundancy.

It is clear from Figures 5 and 6 that the unavailability and redundancy objectives are competing with the LCC objective. However, in most cases, the unavailability and redundancy objectives are non-competing among each other. In some cases, on the other hand, as shown in Figure 6c, some solutions almost form a horizontal line in which the unavailability is reduced while the redundancy remains the same, or even worsens.

Each point in the obtained Pareto-optimal set provides an optimal maintenance solution, in which a balance between the unavailability, redundancy and LCC is achieved. Three representative solutions (S_1, S_2, S_3) are selected from the Pareto-optimal set shown in Figures 5 and 6 are presented as examples of these maintenance solutions. Figure 7 shows (a) the schedules of maintenance application and (b) the history profiles for the cumulative

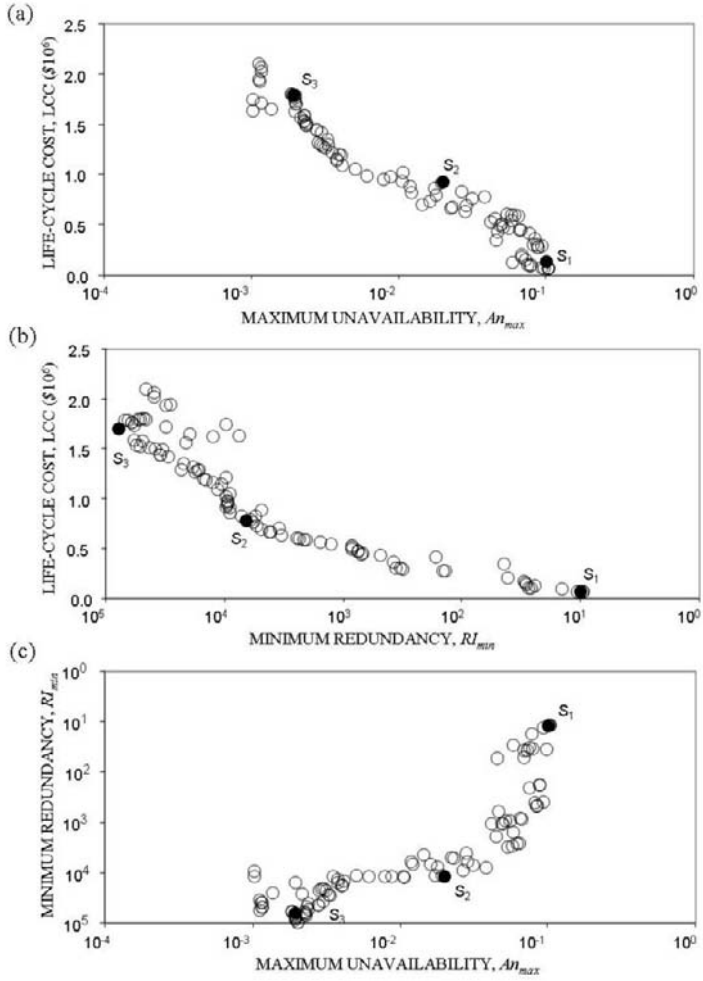


FIGURE 6. Projections of the Pareto-optimal set for Bridge E-17-AH in the bidimensional spaces (adapted from [15])

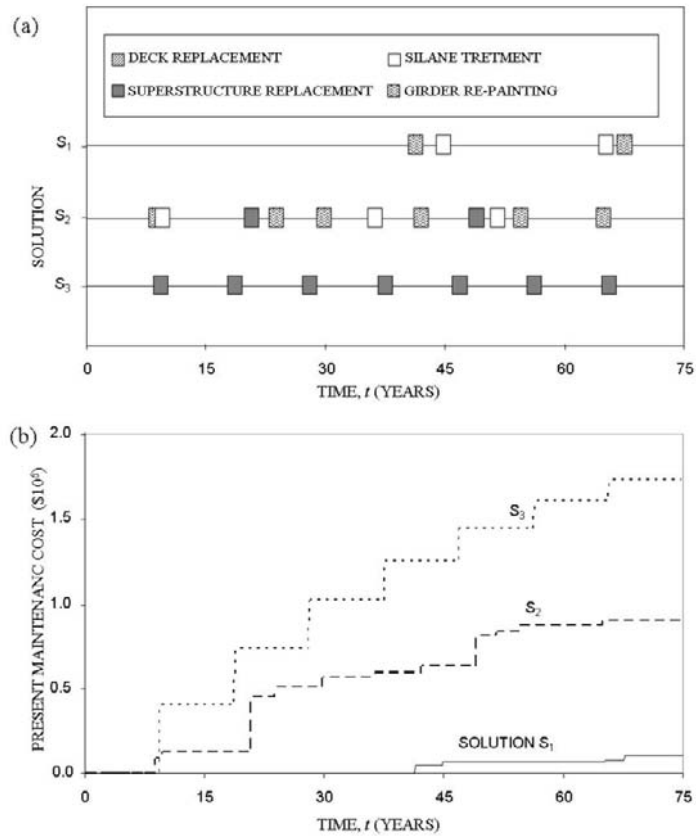


FIGURE 7. (a) Schedules of maintenance application and (b) the history profiles for the cumulative present LCC for the selected solutions (adapted from [15])

present LCC associated with these selected solutions. Also, the history profiles for the unavailability and redundancy are plot in Figures 8 and 9, respectively. Solution S_1 requires no applications of essential maintenance over the lifespan of 75 years and keeps the unavailability below 10^{-1} and redundancy above 10^1 with only two silane treatments (at years 45, and 65) and two girder re-paintings (at years 41, and 68). Solution S_2 requires two superstructure replacements (at years 21, and 49) six girder re-paintings (at years 9, 24, 30, 42, 55, and 65) and three silane treatments (at years 10, 36, and 52) to provide an availability of $10^{-1.77}$ and a redundancy of $10^{3.98}$. The difference in the LCC between Solution S_1 and Solution S_2 is significant.

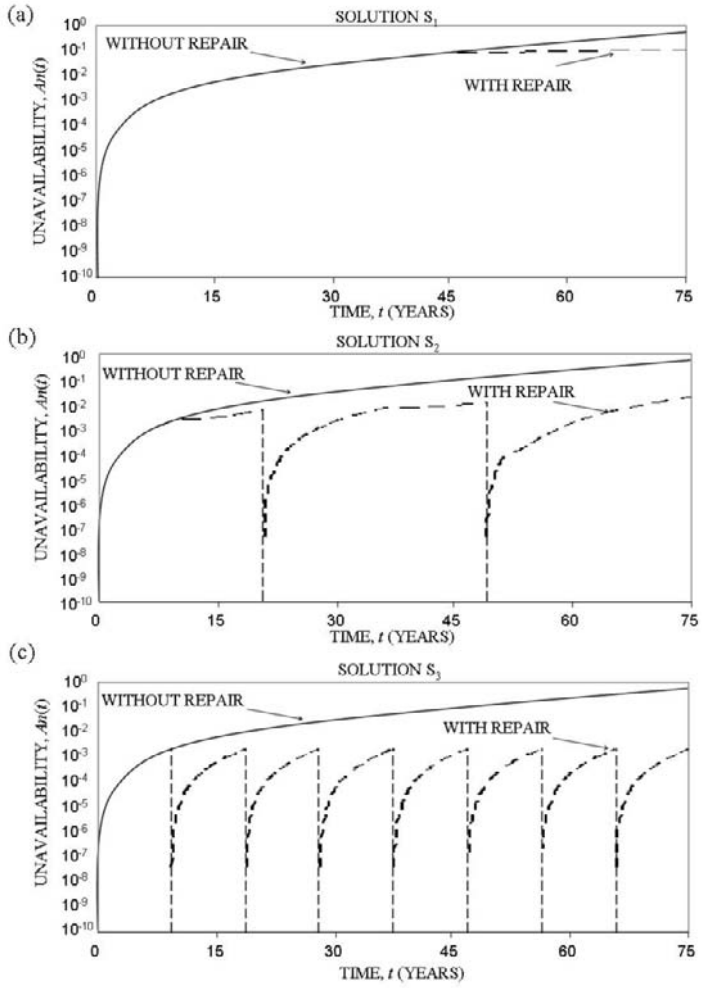


FIGURE 8. History profiles of unavailability for selected solutions (adapted from [15])

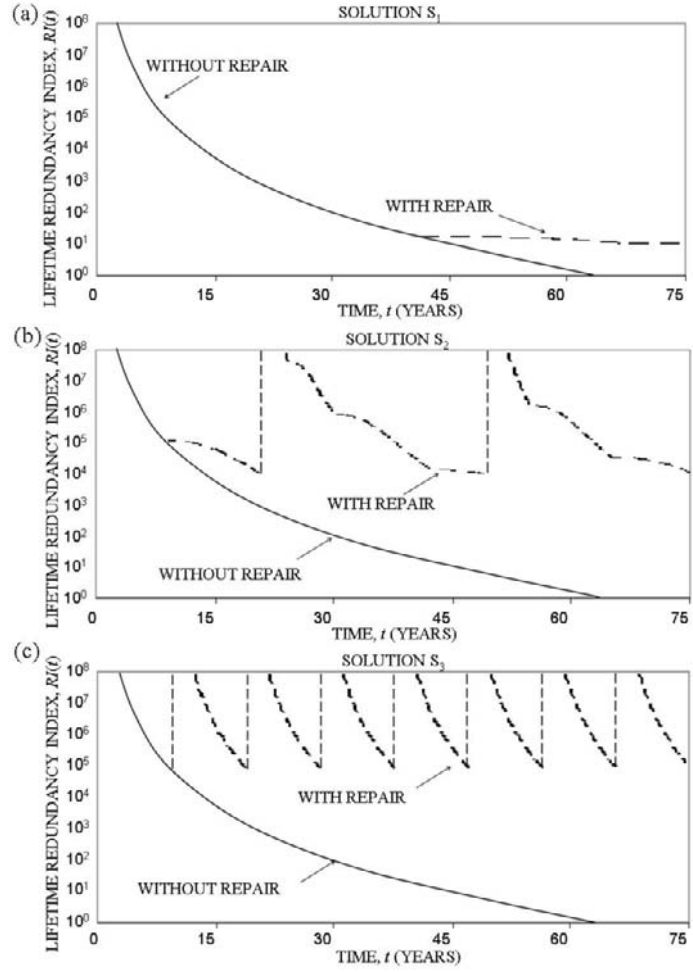


FIGURE 9. History profiles of redundancy for selected solutions (adapted from [15])

Solution S_3 requires seven superstructure replacements (at years 9, 19, 28, 38, 47, 56, and 66) to provide an availability of $10^{-2.71}$ and a redundancy of $10^{4.91}$. Nevertheless, the improvement in unavailability and redundancy compared to solution S_2 requires over eight times the LCC of Solution S_2. Evidently, solving an optimization problem in this nature provides valuable insight regarding the interaction among the different criteria considered and helps decide an optimum maintenance schedule.

5 CONCLUSIONS

In this paper, recent advances in methods of multi-criteria optimization of life-cycle performance of structural systems under uncertainty are reviewed. Two approaches for finding optimum maintenance strategies for deteriorating structural systems through multi-criteria optimization and using genetic algorithms are presented with applications. These approaches use different problem formulations and types of performance indicators. Using an appropriate formulation for the maintenance optimization problem, representative performance measures, and an efficient tool for solving the optimization problem, an economical and effective decision space of optimum maintenance strategies can be obtained, from which decision makers are able to decide their choice based on their preferences, budgets and quality of solutions provided.

Acknowledgments

The support from (a) the National Science Foundation through grants CMS-0638728 and CMS-0639428, (b) the Commonwealth of Pennsylvania, Department of Community and Economic Development, through the Pennsylvania Infrastructure Technology Alliance (PITA), (c) the U.S. Federal Highway Administration Cooperative Agreement Award DTFH61-07-H-00040, and (d) the U.S. Office of Naval Research Contract Number N00014-08-1-0188 is gratefully acknowledged. The opinions and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the sponsoring organizations.

Bibliography

[1] J. M. van Noortwijk and H. E. Klatter. The use of lifetime distributions in bridge maintenance and replacementmodelling. *Computers and Structures*, 82(13–14):1091–1099, 2004.

[2] J. M. van Noortwijk and D. M. Frangopol. Two probabilistic life-cycle maintenance models for deteriorating civilinfrastructures. *Probabilistic Engineering Mechanics*, 19(4):345–359, 2004.

[3] D. M. Frangopol, M. J. Kallen, and J. M. van Noortwijk. Probabilistic models for life-cycle performance of deteriorating structures:review and future directions. *Progress in Structural Engineering and Materials*, 6(4):197–212, 2004.

[4] A. C. Estes and D. M. Frangopol. Repair optimization of highway bridges using system reliability approach. *Journal of Structural Engineering*, 125(7): 766–775, 1999.

[5] M. P. Enright and D. M. Frangopol. Maintenance planning for deteriorating concrete bridges. *Journal of Structural Engineering*, 125(12):1407–1414, 1999.

[6] S-I. Yang, D. M. Frangopol, and L. C. Neves. Optimum maintenance strategy for deteriorating structures based on lifetime functions. *Engineering Structures, Elsevier*, 28(2):196–206, 2006.

[7] S-I. Yang, D. M. Frangopol, Y. Kawakami, and L. C. Neves. The use of lifetime functions in the optimization of interventions on existing bridges considering maintenance and failure costs. *Reliability Engineering & System Safety, Elsevier*, 91(6):698–705, 2006.

[8] D. M. Frangopol and M. Liu. Maintenance and management of civil infrastructure based on condition, safety, optimization, and life-cycle cost. *Structure and Infrastructure Engineering, Taylor & Francis*, 3:29–41, 2007.

[9] M. Liu and D. M. Frangopol. Probabilistic maintenance prioritization for deteriorating bridges using a multiobjective genetic algorithm. In *Proceedings of the Ninth ASCE Specialty Conference on Probabilistic Mechanics and Structural Reliability*, Albuquerque, NM, 2004.

[10] L. C. Neves, D. M. Frangopol, and P. J. Cruz. Probabilistic lifetime-oriented multiobjective optimization of bridge maintenance: Single maintenance type. *Journal of Structural Engineering, ASCE*, 132(6):991–1005, 2006.

[11] A. Petcherdchoo, L. C. Neves, and D. M. Frangopol. Optimizing lifetime condition and reliability of deteriorating structures with emphasis on bridges. *Journal of Structural Engineering*, 134(4):544–552, 2008.

[12] N. M. Okasha and D. M. Frangopol. Time-variant redundancy of structural systems. *Structure and Infrastructure Engineering, Taylor & Francis*, doi:10.1080/15732470802664514, 2009.

[13] K. Deb, A. Pratap, Agrawal S., and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Transaction on Evolutionary Computation, IEEE*, 6(2):182–97, 2002.

[14] N. M. Okasha and D. M. Frangopol. Lifetime-oriented multi-objective optimization of structural maintenance considering system reliability, redundancy and life-cycle cost using ga. *Structural Safety (in press)*, doi:10.1016/j.strusafe.2009.06.005, 2009.

[15] N. M. Okasha and D. M. Frangopol. Lifetime functions for multi-criteria optimization of life-cycle maintenance programs considering availability, redundancy and cost. In *Proceedings of the 13th International Conference on Structural Safety and Reliability, ICOSSAR'09*, volume (in press), Osaka, Japan, 2009.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Model based control at WWTP Westpoort

Hans Korving* – Delft University of Technology, Delft and Witteveen+Bos, Deventer, the Netherlands, and Arie de Niet, Peter Koenders, Remmy Neef – Witteveen+Bos, Deventer, the Netherlands

Abstract. The aeration of the activated sludge tank of wastewater treatment plant (WWTP) Westpoort in Amsterdam (the Netherlands) has been optimised using model based control. Discharge limits for the effluent of the treatment plant require total nitrogen (N_{tot}) concentrations below 10 mg/l. N_{tot} levels are reduced using biological nitrification-denitrification. This process is controlled by aeration which consumes a lot of energy. In order to reduce energy, the nitrification-denitrification process is optimised using a non linear regression model for the ammonium (NH_4) concentration. Simulation results show that the total nitrogen concentration in the effluent can be decreased with a lower oxygen concentration, thus consuming less energy. Both nitrogen removal and energy consumption were reduced with ten percent. Currently, the model based control (MBC) is implemented in the actual process control.

1 INTRODUCTION

Recently, the Dutch water boards signed a long-term agreement with The Ministry of Economic Affairs to improve the energy-efficiency of wastewater treatment plants with at least 2% per year and 30% in ten years time. The energy-efficiency coefficient is roughly the amount of removed waste divided by the net energy consumption. Approximately half the energy consumption of wastewater treatment plants is used for aeration of the activated sludge tanks. Optimisation of the activated sludge process, therefore, is an effective way to increase energy-efficiency.

In order to improve the energy-efficiency at wastewater treatment plant (WWTP) Westpoort, a model based control algorithm for the aeration of the activated sludge process has been designed and implemented at the plant. WWTP Westpoort is a large wastewater treatment facility in Amsterdam (the Netherlands). The plant receives both communal and industrial wastewater of about 400,000 i.e. (inhabitant equivalents) per year and has an inflow of 50,000 m^3 per day. Effluent discharge limits require N_{tot} concentrations below 10 mg/l and P_{tot} (total phosphorous) below 1 mg/l. In

*Witteveen+Bos, P.O. Box 233, 7411 AE, Deventer, the Netherlands; telephone: +31-(0)570 697466, e-mail: j.korving@witteveenbos.nl

four parallel aerated activated sludge tanks, nitrogen and phosphorus are removed biologically.

This paper discusses the optimisation of the nitrification-denitrification process at WWTP Westpoort. In section 2, the principles of biological wastewater treatment are explained. Section 3 describes the development of the model in detail and explains how it is used for control. The results are presented and discussed in Section 4. Finally, the conclusions are summarised in Section 5.

The aim of this paper is to show how model based control (MBC) can be applied in wastewater treatment. As such, the emphasis is on the application not on the theory behind it.

2 BIOLOGICAL WASTEWATER TREATMENT

Nitrogen and phosphorus are removed from wastewater by a mixture of bacteria, also known as activated sludge. Figure 1 presents an outline of the activated sludge process. Some of the bacteria require an oxygen-rich environment to convert ammonium (NH_4) into nitrate (NO_3). Other bacteria convert nitrate to nitrogen gas (N_2) which evaporates. These bacteria prefer, however, a low-oxygen regime. Therefore, the activated sludge tank is partially aerated. This process of nitrification and denitrification is delicate and the effectiveness depends on the amount of O_2 that is added to the water, the water temperature, the inflow of wastewater and the amount of activated sludge.

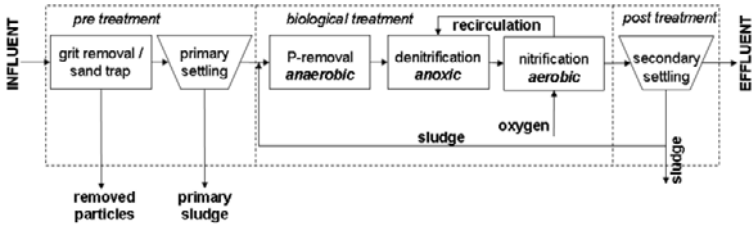


FIGURE 1. Schematic of the activated sludge process

Experiments [1] have shown that, under stationary conditions for temperature and flow, the dependence of N_{tot} (sum of NH_4 and NO_3) of O_2 is close to parabolic. This implies that there is an optimal concentration for O_2. At the optimal concentration, the breakdown of nitrogen is most efficient. Figure 2 shows that a set point lower than the optimal set point for O_2 gives more NH_4, while a higher set point gives more NO_3. However, the trade-off is non linear. At the optimal concentration, the sum of NH_4 and NO_3 is minimal.

Whereas the experiments were done under static conditions, the reality of the activated sludge process is far from static. In practice both flow and

temperature vary. A higher temperature will speed up the conversion of NH_4 and NO_3 by the bacteria and the curves will change such that the optimum moves to a lower O_2 level. A higher flow will increase the levels of NH_4 and more air is required.

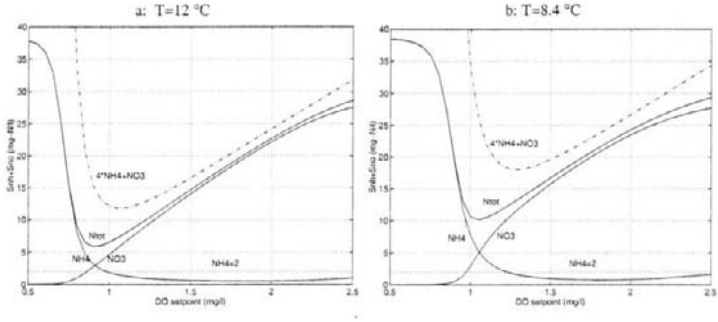


FIGURE 2. Relationship between N_{tot} and O_2 for stationary situations

3 MODEL BASED CONTROL

3.1 Original control strategy

Originally, the choice of the O_2-set point is based on a decision matrix (Table 1) and depends on the measurement of NH_4 and NO_3 in the activated sludge tank. The table shows the control strategy for the tanks at WWTP Westpoort.

High 4 mg NH_4/l	↑↑	↑↑	↑
Acceptable	↑	0	↓
Low 1 mg NH_4/l	0	↓	↓↓
	Low 1 mg NO_3/l	Acceptable	High 6 mg NO_3/l

↑ 1 step up = 0.1 mg O_2/l
↓ 1 step down = 0.1 mg O_2/l

TABLE 1. Decision matrix for O_2 set points

The results of control based on this matrix are good. A yearly average of N_{tot} 6.6 mg/l and P_{tot} 0.7 mg/l is reached which is below discharge limits. However, there is room for improvement, including a more stable process, lower concentrations of N_{tot} and P_{tot}, and lower average O_2 concentrations in the activated sludge tank. Concurrently, the use of energy and chemicals can be reduced.

3.2 Optimisation goals for model based control

The goals of the optimisation of the activated sludge process are:

- better process;
- more stable process;
- cheaper process.

In order to reach these goals, model based control is introduced at WWTP Westpoort. The application of the control algorithm is restricted to dry-weather flow. Flow induced by rainfall simply requires maximal aeration during the event, hence no smart control algorithm is required. About 90% of the time the plant receives dry weather flow.

First,the removal of nitrogen is optimised. With an O_2 set point closer to the optimal value, the nitrogen is removed more efficiently. This leads to lower concentrations in the effluent. With an optimised activated sludge process, the WWTP is able to comply with more strict discharge limits in the future without expensive plant modifications.

Due to daily variation of the flow, NH_4 levels in the activated sludge tank vary. Large oscillations in the concentration diminish removal efficiency. Hence, the second goal is to flatten the peaks in the NH_4 concentration. This can be done by proactive control of the aeration. A model that can predict the increase will start aeration earlier, thus flattening NH_4 peaks. This leads to less varying O_2 set points and less maintenance of the O_2 supply system.

The third goal is a side-effect of the previous two. Due to a more efficient process, less air is needed to remove the same amount of NH_4. Less aeration means less energy consumption, hence lower cost. Due to a more stable process the aeration beds need to be turned on and off less frequently, which increases the lifetime and decreases the maintenance costs of the aeration beds.

3.3 Data analysis

In order to construct the control model, measurement data from one of the tanks of WWTP Westpoort is analysed. The available data comprise: activated sludge temperature, meteorological data, blower set points, logbooks and process data, including NH_4, PO_4, air flow, NO_3, O_2, influent flow and effluent flow.

The temperature of the activated sludge is determined by the air temperature (long term) and the occurrence of rainfall events (short term). During dry weather flow (dwf), the temperature is inversely proportional to the influent flow. However, this variation is smaller than due to rainfall.

The influent flow is divided in a flow through the primary settling tank and a bypass flow. For flows larger than 5,000 m^3/h both parts are highly correlated. The maximum dwf equals 4,500 m^3/h. During dwf, the influent

shows a daily pattern with a global minimum around 5 AM of 1,000 m^3/h, a rapidly increase to a global maximum of 4,000 m^3/h at 12 AM and a local maximum at 8 PM.

Two different control protocols for O_2 set points are applied. The first protocol is based on gradual changes, the second abrupt changes between static minimum and maximum values of set points. The gradual protocol results in less turbulent behaviour and more efficient performance in terms of energy. O_2 concentrations show a strong positive correlation (0.94) with the set points indicating that the aeration protocol can follow the set points very well.

NH_4 concentrations are strongly related to the influent flow. At night, nearly all NH_4 is converted into NO_3 (low influent flow). By day, however, NH_4 levels remain higher after nitrification. The NH_4 concentration also has a high positive correlation with rainfall.

NO_3 concentrations at the end of the denitrification zone remain low indicating that the denitrification process functions properly. NO_3 concentrations at the end of the nitrification zone show a daily pattern similar to the NH_4 concentrations. NO_3 concentrations at both locations show a lower limit which is caused by the recirculation control.

N_{tot} concentrations (sum of NH_4 and NO_3) show a daily pattern with large peaks during rainfall events resulting from an increase of NH_4. These concentrations correspond with the daily pattern of the influent flow. The time lag between influent and N_{tot} equals approximately 3 hours. This is caused by the mixing of influent in the activated sludge tank.

Unfortunately, the theoretical relationship between N_{tot} and O_2 cannot be derived from the measured dataset. First, the dataset does not include all possible situations. Periods with low inflow, e.g. at night, predominantly involve low O_2 set points, whereas high O_2 set points mainly occur during periods with high inflow. Second, the dataset is dominated by situations where the control algorithm operates at the minimum or maximum O_2 set point. Third, the measurements suffer from missing values and signal noise. This significantly reduces the information content of the dataset. As a result, the dataset has limitations for model based optimisation.

In order to overcome the limitations of the measurements, a synthetic dataset is created using a calibrated model of the treatment plant. For this purpose, the most important impacts (active sludge temperature (T) and influent flow(Q)) and the control parameter (O_2 set point) are varied. Variations are based on gradients and ranges which are observed in reality. Consequently, the model results include a proportional combination of all possible situations (except for rainfall events). In addition, the synthetic dataset does not suffer from missing values and signal noise.

3.4 Design of control model

A control model is developed which is based on the (theoretic) relation between O_2 and nitrogen. The model has to describe the process accurate

enough. Otherwise, it cannot be used in a control loop. With a model that fits the current state and accurately predicts the next state of the system the optimal O_2 set point can be determined. The model should reflect the dynamic behaviour and find the optimal O_2 set point under varying conditions.

The state of art IAWQ model for activated sludge processes [2, 3], however, is too complex to be used in real-time control and requires input from laboratory experiments. Therefore, a statistical process model is used that allows for implementation on a PLC (programmable logic controller).

Linear regression models are unsuitable for model based control of the aeration in the activated sludge tank because they cannot describe the complete range of influent flows and active sludge temperatures. Consequently, the possibilities of anticipating changes in these parameters are limited. In addition, the underlying process is non linear and cannot be fully described by a linear model.

Non linear regression models are more appropriate for model based control because they involve more physical relationships. Models predicting NH_4 concentrations give better results than N_{tot} and NO_3. The curve for NH_4 is approximated with a hyperboloid which is based on the following parameters: O_2, temperature and flow.

The basis of the model is the inversely proportional relation between NH_4 and O_2, as shown in Figure 2,

$$f(O_2, \beta) = NH_4 = \frac{1}{O_2}.$$

In addition, two important impacts are included in the model: influent flow and active sludge temperature. Consequently, the regression function becomes,

$$f(O_2, \beta) = NH_4 = \beta_1 \left(\frac{\beta_2(Q/1000) - T + \beta_3}{(O_2 - \beta_4)} \right) + \beta_5,$$

where NH_4 is the ammonium concentration (mg/l), $\beta = (\beta_1, \ldots, \beta_5)^T$ is the vector with unknown parameters, Q is the influent flow (m^3/h), T is the activated sludge temperature and O_2 is the oxygen concentration (mg/l).

Since the regression function is non linear, there exists no explicit solution and an iterative method is needed. In order to estimate the parameter vector β the Gauss-Newton algorithm is used starting with an initial estimation of β. This estimation is improved using a linear approximation of the regression function $f(O_2, \beta)$.

In contrast with the physical relationship between NH_4 and O_2, the regression function has a vertical asymptote for $O_2 \to \beta_4$. The solution to this problem is omitting observations where the O_2 set point is smaller than X_1. This does not result in loss of information because values smaller than X_1 are outside the actual control range of the algorithm. A unique solution

can be found when $X_1 = 1.1$ mg/l, irrespective of the initial estimation of β. The results of the parameter estimation for different O_2 ranges are presented in Table 2.

Model	O_2 range (mg/l)	β_1	β_2	β_3	β_4	β_5	MSE
1	[1.1,4]	0.71	0.09	21.93	-0.44	-0.56	0.48
2	[1.1,5]	0.64	0.11	21.02	-0.56	-0.18	0.38
3	[1.1,6]	0.61	0.13	20.46	-0.62	0.01	0.31
4	[1.1,∞)	0.58	0.15	19.74	-0.68	0.22	0.23

TABLE 2. O_2 ranges and parameter values of non linear models for NH_4

The determination of the O_2 set point is a trade-off between maximisation of treatment performance and minimisation of energy use. The former can be translated into minimisation of the NH_4 concentration in the effluent which requires more aeration, the latter into minimisation of the aeration.

The ideal O_2 set point is located in the bend of the NH_4 curve (Figure 3). The trade-off between the two goals can be described with the angle of the tangent of the curve (α). The goal is to find the point p where the tangent equals α. The corresponding O_2 set point is X and the predicted NH_4 concentration Y.

A larger value for α produces a steeper tangent. The corresponding O_2 set point (X) is smaller and the resulting NH_4 concentration (Y) is larger. A smaller value for α has the opposite effect. This means that larger values for α give priority to minimisation of energy and smaller values for α to minimisation of NH_4.

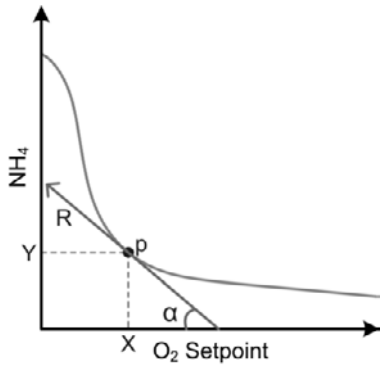


FIGURE 3. Choice of set point for O_2 based on tangent of NH_4 curve

4 RESULTS AND DISCUSSION

4.1 Model based control in SIMBA

The model based control algorithm has been tested using a SIMBA model of WWTP Westpoort. SIMBA is a Matlab-Simulink implementation of the IAWQ model and can be used for dynamic modelling of wastewater treatment plants. Table 3 shows the results of the simulations with the different models for NH_4 in comparison with the original control strategy based on the decision matrix. The results confirm that larger values for α save on aeration (up to 10%) and have considerably lower O_2 set points. The benefit, however, becomes smaller with increasing values of a. In terms of treatment performance $\alpha = 45^o$ gives the best results for N_{tot}. Compared with the decision matrix, performance is increased with 10%. Overall, model 4 gives the best results in terms of aeration and treatment performance. Additionally, treatment performance improves when activated sludge temperatures are higher.

	angle (^{o})	setpoint (mg/l)	Q air (m^3/h)	Q air (%)	N_{tot} (mg/l)	N_{tot} (%)
Decision matrix	-	3.35	154,942	100.0	5.98	100.0
Model 1	60	2.10	137,257	88.6	5.56	93.0
Model 2	60	2.05	137,310	88.6	5.61	93.9
Model 3	60	2.02	139,935	90.3	5.65	94.5
Model 4	60	1.99	136,644	88.2	5.69	95.3
Model 1	45	2.63	143,786	92.8	5.42	90.8
Model 2	45	2.53	142,100	91.7	5.39	90.1
Model 3	45	2.47	141,930	91.6	5.37	89.9
Model 4	45	2.40	140,584	90.7	5.36	89.6
Model 1	30	3.32	151,801	98.0	5.76	96.4
Model 2	30	3.15	149,436	96.4	5.67	94.8
Model 3	30	3.05	149,105	96.2	5.61	93.9
Model 4	30	2.94	148,874	96.1	5.54	92.7

TABLE 3. Simulation results of non linear regression models for NH_4

4.2 Model based control in reality

Currently, the optimisation algorithm is implemented at WWTP Westpoort. As the reality at the plant differs from the SIMBA model, the parameters of the model need to be adapted. The tuning of the parameters is done with a set of rules of a thumb. However, in practice it appears to take some effort and time. Fortunately, the algorithm is robust and not sensitive for sub-optimal parameters. Even then the model based control performs well.

The model uses measurements of temperature and flow for the com-

putation of the O_2 set point. Both measurements are known to be very robust. For both temperature and flow a time-moving average over half an hour is used as model input. This is done for two reasons: response time of the process (about half an hour) and filtering of high-frequent noise in the measurement signal. If one of the measurements fails for more than five minutes, the original decision matrix is used instead of the MBC algorithm.

Even though it is too early to draw final conclusions about the performance of the model based control in practice, some preliminary results can be shown. The MBC compared to the decision matrix gives

- considerably lower O_2 set points (up to a factor of 2);
- more quiet behavior of the O_2 set point;
- higher NH_4 and lower NO_3 concentrations.

A side-effect of the model based control appears to be a decrease in the performance for PO_4. However, the algorithm was not designed to control PO_4. The decreased performance might be caused by the fact that the implementation and the tuning of the parameters took place during summer. At high temperatures, the lower limit for O_2 is determined by the removal of PO_4 instead of N_{tot}. A temporarily decreased angle (30^o instead of 45^o), which leads to a higher level of O_2, is sufficient to maintain the benefits of the MBC and keep the levels of PO_4 within the acceptable range. The results in Table 3 show that with a lower angle the gain in energy and N_{tot} decreases. However, the MBC performs better than the decision matrix. Because high levels of PO_4 occur mainly during high temperatures in summer, it is likely that decreasing temperatures in autumn will allow the angle to be set back to 45^o.

4.3 Further development

A first improvement of the algorithm would be an extension with models for NO_3 and PO_4. Consequently, the choice of the set point for O_2 can be made in a more sophisticated way than in the current model. At least discharge limits for PO_4 can be taken in account.

Second, uncertainty can be introduced in the model. The measurements of flow, temperature, ammonium and nitrate have limited accuracy and sometimes show irregular behaviour. A model that accounts for the inherent uncertainty of measurements could distinguish between noise and signal and react properly to sudden changes.

A third improvement would be an auto-adaptive model that changes the parameters of the model based on observations for NH_4, NO_3 and PO_4. The advantage of an auto-adaptive model is the absence of the time-consuming tuning period. Moreover, an adaptive model can deal more easily with structural changes in influent quality or plant infrastructure.

5 CONCLUSIONS

The objective of this paper is to describe the development and implementation of an optimisation algorithm for the nitrification-denitrification process of an activated sludge tank at WWTP Westpoort. Since the state of art IAWQ model for activated sludge processes is too complex and requires very detailed input, a statistical model is used that allows for easy implementation on site.

The activated sludge process is optimised using a non linear regression model for the NH_4 concentration. This relatively simple model predicts NH_4 based on O_2, T and Q. These parameters represent very robust measurements. The model can approximate the theoretical paraboloid curve which describes the relationship between NH_4 and O_2 accurate enough. The model has several advantages. It is simple, robust and not sensitive to parameter settings.

Simulation results show that N_{tot} concentrations in the effluent can be decreased at lower O_2 set points. These findings are supported by the preliminary results at the plant where the model based control is implemented. Lower set points lead to considerable energy savings. The reduction in both N_{tot} and energy consumption partly depends on the choice of the tangent of the NH_4 curve. With this angle the operator can emphasise either cost reduction or optimal removal of NH_4. Overall, the simulations show that the model based control reduces N_{tot} in the effluent with approximately 10% during dry weather conditions and reduces energy consumption for aeration with 5-10% depending on the angle in the optimisation algorithm.

Acknowledgments

This paper describes the results of a research project, which was carried out in co-operation with Waternet (Amsterdam, The Netherlands). The authors would like to thank Waternet for providing practical knowledge and process data of WWTP Westpoort. They are also grateful to Floris Beltman for his valuable contribution to the research.

Bibliography

[1] S. Weijers. *Modelling, Identification and Control of Activated Sludge Plants for Nitrogen Removal.* PhD thesis, TU Eindhoven, 2000.

[2] M. Henze, C. P. L. Grady, W. Gujer, G. v. R. Marais, and T. Matsuo. Activated sludge model no. 1. In *IAWPRC Scientific and Technical Reports No. 1*, 1987. Londen.

[3] M. Henze, W. Gujer, T. Mino, T. Matsuo, M. C. Wentzel, and Marais G. v. R. Activated sludge model no. 2. In *IAWQ Scientific and Technical Reports No. 3*, 1995. Londen.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Modelling track geometry by a bivariate Gamma wear process, with application to maintenance

Sophie Mercier* – Université de Pau et des Pays de l'Adour, France,
Carolina Meier-Hirmer – SNCF, Paris, France
and Michel Roussignol – Université Paris-Est, Marne-la-Vallée, France

Abstract. This paper discusses the maintenance optimization of a railway track, based on the observation of two dependent randomly increasing deterioration indicators. These two indicators are modelled through a bivariate Gamma process constructed by trivariate reduction. Empirical and maximum likelihood estimators are given for the process parameters and tested on simulated data. The EM algorithm is used to compute the maximum likelihood estimators. A bivariate Gamma process is then fitted to real data of railway track deterioration. Preventive maintenance scheduling is studied, ensuring that the railway track keeps a good quality with a high probability. The results are compared to those based on both indicators taken separately, and also on one single indicator (usually taken for current track maintenance). The results based on the joined information are proved to be safer than the other ones, which shows the interest of the bivariate model.

1 INTRODUCTION

This paper is concerned with the maintenance optimization of a railway track, based on the observation of two dependent randomly increasing deterioration indicators. The railway track is considered as deteriorated when any of these two indicators is beyond a given threshold. The point of the paper is the study of preventive maintenance scheduling, which must ensure that, given some observations provided by inspection, the railway track will remain serviceable until the next maintenance action with a high probability.

Track maintenance is a very expensive task to accomplish. Consequently, it is essential to carry out maintenance actions in an optimal way, while taking into account many parameters: safety and comfort levels to be guaranteed, available logistic means, ... The earlier the deterioration is detected,

*corresponding author: Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques et de leurs Applications – PAU (UMR CNRS 5142), Bâtiment IPRA, Avenue de l'Université – BP 1155, 64013 PAU CEDEX, France; telephone: +33-(0)5 59 40 75 37, fax: +33-(0)5 59 40 75 55, e-mail: sophie.mercier@univ-pau.fr

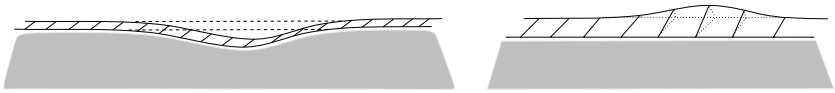


FIGURE 1. Levelling defects

the easier it is to schedule maintenance actions. The objective is therefore to develop a good prediction model.

Deterioration of track geometry is characterized by the development of different representative parameters like, for example, the levelling of the track. Figure 1 shows the defects that are measured by two of these parameters: the longitudinal (NL) and transversal (NT) levelling indicators.

At the SNCF (French National Railways), track inspections are programmed annually on a national level. The interval between two inspections on high speed tracks is currently about two weeks, the inspections are carried out by a modified high-speed train. The collected time series are transformed into indicators that sum up the state of the track over each km. These new indicators are referred to as synthesized Mauzin data. Numeric Mauzin data are available since the opening of the French high-speed lines.

Usually, the synthesized Mauzin indicator of the longitudinal levelling (NL indicator) is used for maintenance issues: thresholds are fixed for this indicator in order to obtain a classification of the track condition and to fix dates for maintenance operations. For example, an intervention should be scheduled before the NL indicator exceeds 0.9.

Based on expert judgements, a Gamma process has been used in [1] both to model the evolution of the NL indicator and to plan maintenance actions. As noted by J.M. van Noortwijk in his recent survey [2], this kind of process is widely used in reliability studies (see also [3], [4] and [5]). Various domains of applications exist, such as civil engineering ([6], [7]), highway engineering [8] or railway engineering [9]. Gamma processes are also used in other domains, such as finance [10] or risk analysis [11]. All these papers use univariate Gamma processes.

In the present case, as the two indicators NL and NT are dependent, the use of a bivariate model is required. For this purpose, different processes might be used, such as Bessel [12] or Lévy processes [13]. In this paper, the approach of F.A. Buijs, J.W. Hall, J.M. van Noortwijk and P.B. Sayers in [6] is used: a specific Lévy process called bivariate Gamma process is considered. This process is constructed from three independent univariate Gamma processes by trivariate reduction, and has univariate Gamma processes as marginal processes.

It is the first time that both NL and NT indicators are used conjointly to predict the optimal dates of maintenance actions. The objective is to analyse

the correlation between the two processes and to determine in what circumstances this bivariate process allows a better prediction of the maintenance times than the current univariate one, based only on the NL indicator.

The paper is organized in the following way: bivariate Gamma processes are introduced in Section 2. Empirical and maximum likelihood estimators for their parameters are provided in Section 3. An EM algorithm is proposed to carry out the maximum likelihood estimation. Both methods are tested on simulated data. Section 4 is devoted to the study of preventive maintenance planning and to the comparison of the results based on the bivariate and on the univariate models. Finally, a bivariate Gamma process is fitted to real data of railway track deterioration in Section 5 and it is shown that the preventive maintenance scheduling based on the two available deterioration indicators are clearly safer than those based on a single one, or on both taken separately.

2 THE BIVARIATE GAMMA PROCESS

Recall that an univariate (homogeneous) Gamma process $(Y_t)_{t\geq 0}$ with parameters $(\alpha, b) \in \mathbb{R}_+^{*2}$ is a process with independent increments such that Y_t is Gamma distributed $\Gamma(\alpha t, b)$ with probability density function (p.d.f.)

$$f_{\alpha t,b}(x) = \frac{b^{\alpha t}}{\Gamma(\alpha t)} x^{\alpha t-1} e^{-bx} \mathbf{1}_{\mathbb{R}_+}(x),$$

$\mathbb{E}(Y_t) = \frac{\alpha t}{b}$, $\mathrm{Var}(Y_t) = \frac{\alpha t}{b^2}$ for all $t > 0$, and $Y_0 \equiv 0$ (see [2] for more details).

Following [6], a bivariate Gamma process $\left(X_t\right)_{t\geq 0} = \left(X_t^{(1)}, X_t^{(2)}\right)_{t\geq 0}$ is constructed by trivariate reduction: starting from three independent univariate Gamma processes $\left(Y_t^{(i)}\right)_{t\geq 0}$ with parameters $(\alpha_i, 1)$ for $i \in \{1, 2, 3\}$ and from $b_1 > 0$, $b_2 > 0$, one defines:

$$X_t^{(1)} = \left(Y_t^{(1)} + Y_t^{(3)}\right)/b_1, \text{ and } X_t^{(2)} = \left(Y_t^{(2)} + Y_t^{(3)}\right)/b_2 \quad \text{for all} \quad t \geq 0.$$

The process $(X_t)_{t\geq 0} = \left(X_t^{(1)}, X_t^{(2)}\right)_{t\geq 0}$ is then a homogeneous process in time with independent increments and it is a Lévy process. The marginal processes of $(X_t)_{t\geq 0}$ are univariate Gamma processes with respective parameters (a_i, b_i), where $a_i = \alpha_i + \alpha_3$ for $i = 1, 2$.

For any bivariate Lévy process, the correlation coefficient ρ_{X_t} of $X_t^{(1)}$ and $X_t^{(2)}$ is known to be independent of t. For a bivariate Gamma process, one obtains:

$$\rho = \rho_{X_t} = \frac{\alpha_3}{\sqrt{a_1 a_2}}$$

and

$$\alpha_1 = a_1 - \rho\sqrt{a_1 a_2}, \quad \alpha_2 = a_2 - \rho\sqrt{a_1 a_2}, \quad \alpha_3 = \rho\sqrt{a_1 a_2}.$$

This entails

$$0 \leq \rho \leq \rho_{\max} = \frac{\min(a_1, a_2)}{\sqrt{a_1 a_2}}. \tag{1}$$

See [14] section XI.3 for results on bivariate Gamma distributions.

This leads to two equivalent parameterizations of a bivariate Gamma process: $(\alpha_1, \alpha_2, \alpha_3, b_1, b_2)$ and $(a_1, a_2, b_1, b_2, \rho)$.

With the parameterization $(\alpha_1, \alpha_2, \alpha_3, b_1, b_2)$, the joint p.d.f. of X_t is:

$$\begin{aligned} & g_t(x_1, x_2) \\ & = b_1 b_2 \int_0^{\min(b_1 x_1, b_2 x_2)} f_{\alpha_1 t,1}(b_1 x_1 - x_3)\, f_{\alpha_2 t,1}(b_2 x_2 - x_3)\, f_{\alpha_3 t,1}(x_3)\ dx_3, \\ & = \frac{b_1 b_2 e^{-b_1 x_1 - b_2 x_2}}{\Gamma(\alpha_1 t)\,\Gamma(\alpha_2 t)\,\Gamma(\alpha_3 t)} \times \cdots \\ & \times \int_0^{\min(b_1 x_1, b_2 x_2)} (b_1 x_1 - x_3)^{\alpha_1 t - 1} (b_2 x_2 - x_3)^{\alpha_2 t - 1} x_3^{\alpha_3 t - 1} e^{-x_3}\ dx_3. \end{aligned} \tag{2}$$

3 PARAMETER ESTIMATION

The data used for the parameter estimation are values of the process increments for non overlapping time intervals on a single trajectory, and also on different independent trajectories. The data can then be represented as $\left(\Delta t_j, \Delta X_j^{(1)}(\omega), \Delta X_j^{(2)}(\omega)\right)_{1 \le j \le n}$ where $\Delta t_j = t_j - s_j$ stands for a time increment and $\Delta X_j^{(i)} = X_{t_j}^{(i)} - X_{s_j}^{(i)}$ for the associated i-th marginal increment $(i = 1, 2)$. For different j, the random vectors $\left(\Delta X_j^{(1)}, \Delta X_j^{(2)}\right)$ are independent, but not identically distributed. The random variable $\Delta X_j^{(i)}$ $(i = 1, 2)$ is Gamma distributed with parameters $(a_i\, \Delta t_j, b_i)$. The joint p.d.f. of the random vector $\left(\Delta X_j^{(1)}, \Delta X_j^{(2)}\right)$ is equal to $g_{\Delta t_j}(.,.)$, with Δt_j substituted for t in (2). In the same way as for parameter estimation of a (univariate) Gamma process, both empirical and maximum likelihood methods are possible in the bivariate case.

3.1 Empirical estimators

Using $\mathbb{E}\left(\Delta X_j^{(i)}\right) = \frac{a_i}{b_i}\, \Delta t_j$ and $\mathrm{Var}\left(\Delta X_j^{(i)}\right) = \frac{a_i}{b_i^2}\, \Delta t_j$ for $i = 1, 2$ and for all j, empirical estimators $(\hat{a}_1, \hat{b}_1, \hat{a}_2, \hat{b}_2)$ of (a_1, b_1, a_2, b_2) are given in [7] and [15], with:

$$\frac{\hat{a}_i}{\hat{b}_i} = \frac{\sum_{j=1}^n \Delta X_j^{(i)}}{t_n} \quad \text{and} \quad \frac{\hat{a}_i}{\hat{b}_i^2} = \frac{\sum_{j=1}^n \left(\Delta X_j^{(i)} - \frac{\hat{a}_i}{\hat{b}_i} \Delta t_j\right)^2}{t_n - \frac{1}{t_n} \sum_{j=1}^n (\Delta t_j)^2}, \tag{3}$$

where we set $t_n = \sum_{j=1}^n \Delta t_j$. Using

$$\mathrm{Cov}\left(\Delta X_j^{(1)}, \Delta X_j^{(2)}\right) = \rho \frac{\sqrt{a_1 a_2}}{b_1 b_2} \Delta t_j,$$

a similar estimator $\hat{\rho}$ may be given for ρ, with:

$$\hat{\rho}\frac{\sqrt{\hat{a}_1\hat{a}_2}}{\hat{b}_1\hat{b}_2} = \frac{\sum_{j=1}^{n}\left(\Delta X_j^{(1)} - \frac{\hat{a}_1}{\hat{b}_1}\Delta t_j\right)\left(\Delta X_j^{(2)} - \frac{\hat{a}_2}{\hat{b}_2}\Delta t_j\right)}{t_n - \frac{1}{t_n}\sum_{j=1}^{n}\left(\Delta t_j\right)^2}. \tag{4}$$

These estimators satisfy:

$$\mathbb{E}\left(\frac{\hat{a}_i}{\hat{b}_i}\right) = \frac{a_i}{b_i}, \quad \mathbb{E}\left(\frac{\hat{a}_i}{\hat{b}_i^2}\right) = \frac{a_i}{b_i^2}, \quad \mathbb{E}\left(\hat{\rho}\frac{\sqrt{\hat{a}_1\hat{a}_2}}{\hat{b}_1\hat{b}_2}\right) = \rho\frac{\sqrt{a_1a_2}}{b_1b_2}.$$

If the time increments Δt_j are equal, these estimators coïncides with the usual empirical estimators in the case of i.i.d. random variables.

3.2 Maximum likelihood estimators

The parameter estimation of an univariate Gamma process is usually done by maximizing the likelihood function (see e.g. [1]). With this method, estimators $\bar{a}_i$ and $\bar{b}_i$ $(i = 1, 2)$ of the marginal parameters are computed by solving the equations:

$$\frac{\bar{a}_i}{\bar{b}_i} = \frac{\sum_{j=1}^{n}\Delta X_j^{(i)}}{\sum_{j=1}^{n}\Delta t_j} \quad \text{and}$$

$$\left(\sum_{j=1}^{n}\Delta t_j\right) \times \ln\left(\bar{a}_i\frac{\sum_{j=1}^{n}\Delta t_j}{\sum_{j=1}^{n}\Delta X_j^{(i)}}\right) + \sum_{j=1}^{n}\Delta t_j\left[\ln\left(\Delta X_j^{(i)}\right) - \psi\left(\bar{a}_i\ \Delta t_j\right)\right] = 0,$$

where

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)},\ \Gamma(x) = \int_0^{\infty} e^{-u}u^{x-1}du$$

for all $x > 0$ (ψ is the Digamma function).

In order to estimate all the parameters of the bivariate process $(\alpha_1, \alpha_2, \alpha_3, b_1, b_2)$ (which are here prefered to $(a_1, b_1, a_2, b_2, \rho)$), the likelihood function associated with the data $\left(\Delta t_j, \Delta X_j^{(1)}, \Delta X_j^{(2)}\right)_{1\leq j\leq n}$ can be written as $\mathcal{L}(\alpha_1, \alpha_2, \alpha_3, b_1, b_2) = \prod_{j=1}^{n} g_{\Delta t_j}(\Delta X_j^{(1)}, \Delta X_j^{(2)})$. However, because of the expression of the function $g_t(.,.)$, it seems complicated to optimize this likelihood function directly. An EM algorithm (see [16]) is then used, considering $\left(\Delta Y_j^{(3)} = Y_{t_j}^{(3)} - Y_{s_j}^{(3)}\right)_{1\leq j\leq n}$ as hidden data. This procedure is still too complicated to estimate all the five parameters and does not work numerically. So, the procedure is restricted to the three parameters $(\alpha_1, \alpha_2, \alpha_3)$. For the parameters b_1, b_2, the values $\left(\bar{b}_1, \bar{b}_2\right)$ computed using the maximum likelihood method for each univariate marginal process are taken.

In order to simplify the expressions, the values of the data $\left(\Delta t_j,\ \Delta X_j^{(1)},\ \Delta X_j^{(2)},\ \Delta Y_j^{(3)}\right)_{1\leq j\leq n}$ are denoted by $\left(t_j, x_j^{(1)}, x_j^{(2)}, y_j^{(3)}\right)_{1\leq j\leq n}$, the associated n-dimensional random vectors by $\left(\overline{X}^{(1)}, \overline{X}^{(2)}, \overline{Y}^{(3)}\right)$ and the associated n-dimensional data vectors by $\left(\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}^{(3)}\right)$.

The joint p.d.f. of the random vector $\left(X_t^{(1)}, X_t^{(2)}, Y_t^{(3)}\right)$ is equal to:

$$b_1 b_2 f_{\alpha_1 t,1}\left(b_1 x_1 - y_3\right) f_{\alpha_2 t,1}\left(b_2 x_2 - y_3\right) f_{\alpha_3 t,1}\left(y_3\right) = \frac{b_1 b_2 e^{-(b_1 x_1 + b_2 x_2)}}{\Gamma\left(\alpha_1 t\right)\Gamma\left(\alpha_2 t\right)\Gamma\left(\alpha_3 t\right)} \left(b_1 x_1 - y_3\right)^{\alpha_1 t - 1} \left(b_2 x_2 - y_3\right)^{\alpha_2 t - 1} y_3^{\alpha_3 t - 1} e^{y_3},$$

with $0 \leq y_3 \leq \min\left(b_1 x_1, b_2 x_2\right)$, $x_1 > 0$ and $x_2 > 0$. Then, the log-likelihood function $Q\left(\bar{x}^{(1)}, \bar{x}^{(2)}, \bar{y}^{(3)}\right)$ associated with the complete data $\left(\overline{x}^{(1)}, \overline{x}^{(2)}, \overline{y}^{(3)}\right)$ is derived:

$$\begin{aligned} Q\left(\bar{x}^{(1)}, \bar{x}^{(2)}, \bar{y}^{(3)}\right) &= n\left(\ln\left(b_1\right) + \ln\left(b_2\right)\right) - \cdots \\ &\sum_{j=1}^{n}\left(\ln\Gamma\left(\alpha_1 t_j\right) + \ln\Gamma\left(\alpha_2 t_j\right) + \ln\Gamma\left(\alpha_3 t_j\right)\right) - b_1 \sum_{j=1}^{n} x_j^{(1)} - \cdots \\ &b_2 \sum_{j=1}^{n} x_j^{(2)} + \sum_{j=1}^{n}\Big\{\left(\alpha_1 t_j - 1\right)\ln\left(b_1 x_j^{(1)} - y_j^{(3)}\right) + \cdots \\ &\left(\alpha_2 t_j - 1\right)\ln\left(b_2 x_j^{(2)} - y_j^{(3)}\right) + \left(\alpha_3 t_j - 1\right)\ln\left(y_j^{(3)}\right) + y_j^{(3)}\Big\}. \end{aligned}$$

For the EM algorithm, the conditional log-likelihood of the complete data given the observed data is needed:

$$\begin{aligned} &\mathbb{E}\left(Q\left(\bar{X}^{(1)}, \bar{X}^{(2)}, \bar{Y}^{(3)}\right) | \bar{X}^{(1)} = \bar{x}^{(1)}, \bar{X}^{(2)} = \bar{x}^{(2)}\right) \\ &= n\left(\ln\left(b_1\right) + \ln\left(b_2\right)\right) - b_1 \sum_{j=1}^{n} x_j^{(1)} - b_2 \sum_{j=1}^{n} x_j^{(2)} + \cdots \\ &\sum_{j=1}^{n} \Big\{\left(\left(\alpha_1 t_j - 1\right)\mathbb{E}\left(\ln\left(b_1 x_j^{(1)} - \Delta Y_j^{(3)}\right) | \Delta X_j^{(1)} = x_j^{(1)}, \Delta X_j^{(2)} = x_j^{(2)}\right)\right. \\ &\quad + \left(\alpha_2 t_j - 1\right)\mathbb{E}\left(\ln\left(b_2 x_j^{(2)} - \Delta Y_j^{(3)}\right) | \Delta X_j^{(1)} = x_j^{(1)}, \Delta X_j^{(2)} = x_j^{(2)}\right) \\ &\quad + \left(\alpha_3 t_j - 1\right)\mathbb{E}\left(\ln\left(\Delta Y_j^{(3)}\right) | \Delta X_j^{(1)} = x_j^{(1)}, \Delta X_j^{(2)} = x_j^{(2)}\right) \\ &\quad \left. + \mathbb{E}\left(Y_j^{(3)} | \Delta X_j^{(1)} = x_j^{(1)}, \Delta X_j^{(2)} = x_j^{(2)}\right)\Big\} \\ &- \sum_{j=1}^{n}\left(\ln\Gamma\left(\alpha_1 t_j\right) + \ln\Gamma\left(\alpha_2 t_j\right) + \ln\Gamma\left(\alpha_3 t_j\right)\right). \end{aligned} \tag{5}$$

Finally, the conditional density function of $Y_t^{(3)}$ given $X_t^{(1)} = x_1, X_t^{(2)} = x_2$ is equal to:

$$\begin{aligned} &\frac{f_{\alpha_1 t,1}\left(b_1 x_1 - y_3\right) f_{\alpha_2 t,1}\left(b_2 x_2 - y_3\right) f_{\alpha_3 t,1}\left(y_3\right)}{\int_0^{\min(b_1 x_1, b_2 x_2)} f_{\alpha_1 t,1}\left(b_1 x_1 - x_3\right) f_{\alpha_2 t,1}\left(b_2 x_2 - x_3\right) f_{\alpha_3 t,1}\left(x_3\right)\ dx_3} \\ &= \frac{\left(b_1 x_1 - y_3\right)^{\alpha_1 t - 1}\left(b_2 x_2 - y_3\right)^{\alpha_2 t - 1} y_3^{\alpha_3 t - 1} e^{y_3}}{\int_0^{\min(b_1 x_1, b_2 x_2)}\left(b_1 x_1 - x_3\right)^{\alpha_1 t - 1}\left(b_2 x_2 - x_3\right)^{\alpha_2 t - 1} x_3^{\alpha_3 t - 1} e^{x_3}\ dx_3}, \end{aligned}$$

where $0 \leq y_3 \leq \min(b_1 x_1, b_2 x_2)$, $x_1 > 0$ and $x_2 > 0$.

Step k of the EM algorithm consists of computing new parameter values $(\alpha_1^{(k+1)}, \alpha_2^{(k+1)}, \alpha_3^{(k+1)})$ given the current values $(\alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)})$ in two stages:

- stage 1: compute the conditional expectations in (5) using the current set $(\alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)})$ of parameters, with:

$$
\begin{aligned}
f_1\left(j, \alpha_1^k, \alpha_2^{(k)}, \alpha_3^{(k)}\right) &= \mathbb{E}\left(\ln\left(\bar{b}_1 \bar{x}_j^{(1)} - \bar{Y}_j^{(3)}\right) | \bar{X}^{(1)} = \bar{x}_j^{(1)}, \bar{X}^{(2)} = \bar{x}_j^{(2)}\right), \\
f_2\left(j, \alpha_1^k, \alpha_2^{(k)}, \alpha_3^{(k)}\right) &= \mathbb{E}\left(\ln\left(\bar{b}_2 \bar{x}_j^{(2)} - \bar{Y}_j^{(3)}\right) | \bar{X}^{(1)} = \bar{x}_j^{(1)}, \bar{X}^{(2)} = \bar{x}_j^{(2)}\right), \\
f_3\left(j, \alpha_1^k, \alpha_2^{(k)}, \alpha_3^{(k)}\right) &= \mathbb{E}\left(\ln\left(\bar{Y}_j^{(3)}\right) | \bar{X}^{(1)} = \bar{x}_j^{(1)}, \bar{X}^{(2)} = \bar{x}_j^{(2)}\right), \\
h\left(\alpha_1^k, \alpha_2^{(k)}, \alpha_3^{(k)}\right) &= \sum_{j=1}^{n} \mathbb{E}\left(\bar{Y}_j^{(3)} | \bar{X}^{(1)} = \bar{x}_j^{(1)}, \bar{X}^{(2)} = \bar{x}_j^{(2)}\right).
\end{aligned}
$$

- stage 2: take for $(\alpha_1^{(k+1)}, \alpha_2^{(k+1)}, \alpha_3^{(k+1)})$ the values of $(\alpha_1, \alpha_2, \alpha_3)$ that maximize (5), which here becomes:

$$
\begin{aligned}
&g\big(\alpha_1, \alpha_2, \alpha_3, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) \\
&= n\left(\ln\left(\bar{b}_1\right) + \ln\left(\bar{b}_2\right)\right) - \bar{b}_1 \sum_{j=1}^{n} x_j^{(1)} - \bar{b}_2 \sum_{j=1}^{n} x_j^{(2)} \\
&+ \sum_{j=1}^{n} \Big\{ (\alpha_1 t_j - 1) f_1\big(j, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) + (\alpha_2 t_j - 1) f_2\big(j, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) \\
&\qquad + (\alpha_3 t_j - 1) f_3\big(j, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) \Big\} \\
&- \sum_{j=1}^{n} \left(\ln\Gamma(\alpha_1 t_j) + \ln\Gamma(\alpha_2 t_j) + \ln\Gamma(\alpha_3 t_j)\right) + h\big(\alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big).
\end{aligned}
$$

The maximization in stage 2 is done by solving the following equation with respect to α_i:

$$
\begin{aligned}
&\frac{\partial g\big(\alpha_1, \alpha_2, \alpha_3, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big)}{\partial \alpha_i} = \\
&\qquad \sum_{j=1}^{n} t_j f_i\big(j, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) - \sum_{j=1}^{n} t_j \psi(\alpha_i t_j) = 0 \quad (6)
\end{aligned}
$$

for $i = 1, 2, 3$.

In the same way, it is possible to take the values $\left(\bar{a}_1, \bar{a}_2, \bar{b}_1, \bar{b}_2\right)$ obtained by maximum likelihood estimation on the univariate marginal processes

for (a_1, a_2, b_1, b_2) and to estimate only the last parameter α_3 by the EM algorithm. In that case, $\alpha_3^{(k+1)}$ is the solution of the equation:

$$\sum_{j=1}^{n} t_j \Big\{ f_3\big(j, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) - f_1\big(j, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) - \cdots$$
$$f_2\big(j, \alpha_1^{(k)}, \alpha_2^{(k)}, \alpha_3^{(k)}\big) \Big\} - \sum_{j=1}^{n} t_j \Big\{ \psi\left(\alpha_3 t_j\right) - \psi\left((\bar{a}_1 - \alpha_3) t_j\right) - \cdots$$
$$\psi\left((\bar{a}_2 - \alpha_3) t_j\right) \Big\} = 0.$$

3.3 Tests on simulated data

500 time increments $(t_j)_{1 \le j \le 500}$ are randomly chosen similar to the data of track deterioration (the proposed methods will be used on these data in Section 5). Then, 500 values of a bivariate Gamma process are simulated corresponding to these time increments and with parameters $a_1 = 0.33, a_2 = 0.035, b_1 = 13.5, b_2 = 20$ and $\rho = 0.5296$. These parameter values have the same order of magnitude than those observed for track deterioration studied in Section 5. Three series of 500 data points are simulated independently. Results of parameters estimation are given in Tables 1, 2 and 3, each corresponding to a series of data. In these tables, one can find: the true values in column 2, the empirical estimators in column 3, the univariate maximum likelihood estimators of a_1, b_1, a_2, b_2 in column 4, the EM estimator of the three parameters a_1, a_2, ρ in column 5, using the parameters $\bar{b}_1, \bar{b}_2$ previously estimated by the univariate maximum likelihood method (from column 4), and the second EM estimator of the parameter ρ in column 6, using the estimated parameters $\bar{a}_1, \bar{b}_1, \bar{a}_2, \bar{b}_2$ from column 4.

The initial values for the EM algorithm are different for the three tables. For Table 1, the EM algorithm has been initiated with $\alpha_1^{(0)} = \alpha_2^{(0)} = 0.05$ and $\alpha_3^{(0)} = 0.15$ ($a_1^{(0)} = a_2^{(0)} = 0.1$ and $\rho^{(0)} = 0.75$). For Tables 2 and 3, $\alpha_1^{(0)} = \alpha_2^{(0)} = \alpha_3^{(0)} = 0.01$, and $\alpha_1^{(0)} = 0.02$, $\alpha_2^{(0)} = 0.01$, $\alpha_3^{(0)} = 0.05$ were respectively taken.

Looking at the development of $a_i^{(k)}$ and $\rho^{(k)}$ along the different steps of the EM algorithm, one may note that the parameters $a_i^{(k)}$ stabilize more quickly than the parameter $\rho^{(k)}$ (about 5 iterations for $a_i^{(k)}$ and between 20 and 30 iterations for $\rho^{(k)}$).

The conclusion of this section is that estimation of parameters (a_i, b_i) by empirical and maximum likelihood methods give satisfactory results, with a slight preference to maximum likelihood. Empirical estimators of ρ have a good order of magnitude, but are sometimes not precise enough. Estimators of ρ obtained by EM are always reasonable. The estimation of the three parameters $(\alpha_1, \alpha_2, \alpha_3)$ (column EM1) seems to give slightly better results

	True values	Empirical estimators	Univariate max likelihood	EM algorithm EM1	EM2
a_1	0.0330	0.0348	0.0342	0.0347	–
b_1	13.5	14.38	14.14	–	–
a_2	0.0350	0.0362	0.0357	0.0354	–
b_2	20	20.58	20.25	–	–
ρ	0.5296	0.5637	–	0.5231	0.5214

TABLE 1. Results for the first series of data.

	True values	Empirical estimators	Univariate max likelihood	EM algorithm EM1	EM2
a_1	0.0330	0.0315	0.0326	0.0328	–
b_1	13.5	12.80	13.16	–	–
a_2	0.0350	0.0357	0.0361	0.0365	–
b_2	20	20.25	20.54	–	–
ρ	0.5296	0.5750	–	0.5272	0.5257

TABLE 2. Results for the second series of data.

than those of the estimation of the parameter α_3 alone (column EM2). The results obtained by the EM algorithm for parameters a_i (column EM1) are good, with a quality quite similar to those obtained by univariate maximum likelihood estimation. Finally, the EM algorithm does not seem sensitive to initial values, at least if the initial value of α_3 is not too small.

4 PREVENTIVE MAINTENANCE PLANNING

A bivariate Gamma process $X_t = \left(X_t^{(1)}, X_t^{(2)}\right)$ is now used to model the development of two deterioration indicators of a system. We assume that there exist (corrective) thresholds s_i $(i = 1, 2)$ for each indicator, above which the system is considered to be deteriorated. The system is not continuously monitored but only inspected at time intervals, with a perfect observation of the deterioration level. When one (or both) indicator(s) is observed to be beyond its corrective threshold, an instantaneous maintenance action is undertaken, which brings the system back to a better state, not necessarily as good as new. When both indicators are observed to be below their corrective thresholds or after a maintenance action, a new inspection is planned. The time to next inspection (τ) must ensure with a high probability that neither $X_t^{(1)}$ nor $X_t^{(2)}$ go beyond their corrective thresholds s_i before the next inspection.

Let $(x_1, x_2) \in [0, s_1[\times[0, s_2[$ be the observed deterioration level at some inspection time, say at time $t = 0$ with no restriction. (If $x_1 \geq s_1$ or $x_2 \geq s_2$, a maintenance action is immediately undertaken).

For $i = 1, 2$, let $T^{(i)}$ be the hitting time of the threshold s_i for the marginal process $\left(X_t^{(i)}\right)_{t\geq 0}$, with $T^{(i)} = \inf\left(t > 0 : X_t^{(i)} \geq s_i\right)$. Also, let $\varepsilon \in]0, 1[$ be some confidence level.

	True values	Empirical estimators	Univariate max likelihood	EM algorithm EM1	EM algorithm EM2
a_1	0.0330	0.0297	0.0340	0.0343	–
b_1	13.5	11.71	13.43	–	–
a_2	0.0350	0.0340	0.0385	0.0389	–
b_2	20	18.79	21.28	–	–
ρ	0.5296	0.5645	–	0.5060	0.5027

TABLE 3. Results for the third series of data.

	a_2	b_2	x_1	x_2	$\rho_{\max}$	$\tau^{(1)}$	$\tau^{(2)}$	τ^U	$\tau^B(\rho_{\max})$
case 1	0.03	30	0.2	0.2	1	341.12	558.31	341.12	341.12
case 2	0.04	20	0.4	0.2	0.866	237.33	255.84	237.33	229.91

TABLE 4. Two different combinations of values for a_2, b_2, x_1 and x_2, and the resulting $\rho_{\max}$, $\tau^{(1)}$, $\tau^{(2)}$, τ^U and $\tau^B(\rho_{\max})$.

Different points of view are possible: in the first case, $\tau^{(i)}$, $i = 1, 2$ is the time to next inspection associated to the marginal process $\left(X_t^{(i)}\right)_{t\geq 0}$, with

$$\tau^{(i)} = \max\left(\tau \geq 0 \text{ such that } \mathbb{P}_{x_i}\left(T^{(i)} > \tau\right) \geq 1-\varepsilon\right),$$

where $\mathbb{P}_{x_i}$ stands for the conditional probability given $X_0^{(i)} = x_i$. One then gets: $\mathbb{P}_{x_i}\left(T^{(i)} > \tau^{(i)}\right) = 1-\varepsilon$.

Without a bivariate model, a natural time to next inspection for the system is:

$$\begin{aligned}\tau^U &= \max\left(\tau \geq 0 \text{ s.t. } \mathbb{P}_{x_1}\left(T^{(1)} > \tau\right) \geq 1-\varepsilon \text{ and } \mathbb{P}_{x_2}\left(T^{(2)} > \tau\right) \geq 1-\varepsilon\right),\\ &= \min\left(\tau^{(1)}, \tau^{(2)}\right).\end{aligned}$$

Using a bivariate Gamma process, the time to next inspection becomes:

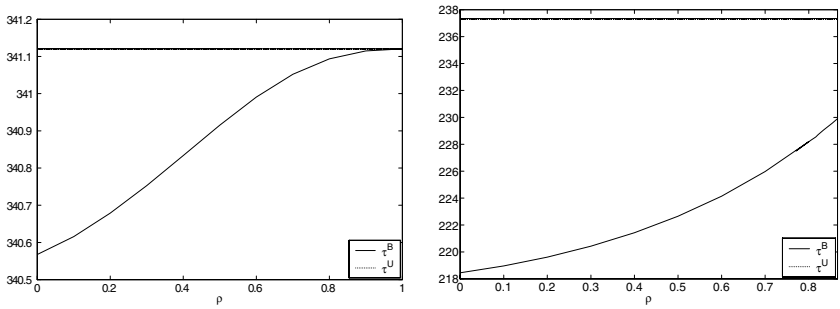
$$\tau^B = \max\left(\tau \geq 0 \text{ such that } \mathbb{P}_{(x_1,x_2)}\left(T^{(1)} > \tau, T^{(2)} > \tau\right) \geq 1-\varepsilon\right).$$

The goal is to compare τ^U and τ^B, and more generally, to understand the influence of the dependence between both components on τ^B. Using

$$\mathbb{P}_{x_i}\left(T^{(i)} > t\right) = \mathbb{P}_{x_i}\left(X_t^{(i)} < s_i\right) = \mathbb{P}_0\left(X_t^{(i)} < s_i - x_i\right) = F_{a_i t, b_i}\left(s_i - x_i\right),$$

where $F_{a_i t, b_i}(x)$ is the cumulative distribution function of the distribution $\Gamma(a_i t, b_i)$, the real $\tau^{(i)}$ is computed by solving the equation $F_{a_i \tau^{(i)}, b_i}(s_i - x_i) = 1-\varepsilon$, for $i = 1, 2$, and $\tau^U = \min\left(\tau^{(1)}, \tau^{(2)}\right)$ is derived. Similarly,

$$\begin{aligned}\mathbb{P}_{(x_1,x_2)}\left(T^{(1)} > t, T^{(2)} > t\right) &= \mathbb{P}_{(0,0)}\left(X_t^{(1)} < s_1 - x_1, X_t^{(2)} < s_2 - x_2\right),\\ &= \int_0^{s_1-x_1}\int_0^{s_2-x_2} g_t(y_1, y_2)\ dy_1\ dy_2,\\ &\equiv G_t\left(s_1 - x_1, s_2 - x_2\right),\end{aligned}$$



case 1 case 4

FIGURE 2. τ^B with respect to ρ and τ^U, for the four cases of Table 4

where g_t is the p.d.f. of X_t (see (2)). This provides τ^B by solving $G_{\tau^B}(s_1 - x_1, s_2 - x_2) = 1 - \varepsilon$.

With $a_1 = 0.03$, $b_1 = 20$, $\epsilon = 0.5$ and $s_1 = s_2 = 1$, and different values for a_2, b_2, x_1 and x_2, Table 4 gives the corresponding values for $\rho_{\max}$ (as provided by (1)) and the resulting $\tau^{(1)}$, $\tau^{(2)}$, τ^U and $\tau^B(\rho_{\max})$. The value of τ^B is plotted with respect to ρ in the Figures 2 for the two different cases of Table 4, and the corresponding value of τ^U is indicated.

In both figures, one can observe that with all other parameters fixed, the bivariate preventive time τ^B is an increasing function of ρ, such that $\tau^B \leq \tau^U$. Also, both $\tau^B = \tau^U$ and $\tau^B < \tau^U$ are possible. The theoretical proof of such results is not provided here because of the reduced size of the present paper, but will be provided in a forthcomming one.

In conclusion to this section, one can see that using a bivariate model instead of two separate univariate models generally shortens the time to next inspection ($\tau^B \leq \tau^U$). This means that taking into account the dependence between both components provides safer results. Also, the optimal time to next inspection is increasing with dependence (τ^B increases with ρ), which implies that the error made when considering separate models (τ^U) is all the more important the less the components are dependent. This also implies that the safest attitude, in case of an unkown correlation, is to consider both components as independent and chose $\tau = \tau^\perp$, where

$$\tau^\perp = \max\big(\tau \geq 0 \text{ such that } \mathbb{P}_{x_1}\big(T^{(1)} > \tau\big)\mathbb{P}_{x_2}\big(T^{(2)} > \tau\big) \geq 1 - \varepsilon\big).$$

5 APPLICATION TO TRACK MAINTENANCE

A bivariate Gamma process is now used to model the evolution of the two track indicators NL and NT (see the Introduction) and times to next inspection are computed, as described in the previous section.

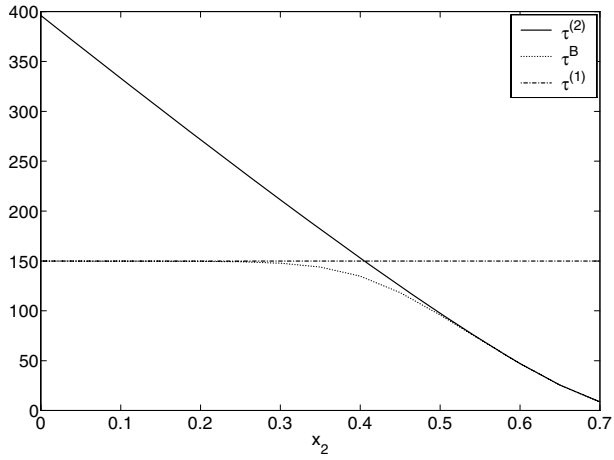


FIGURE 3. $\tau^{(1)}$, $\tau^{(2)}$ and τ^B with respect to x_2 with $x_1 = 0.4$

Using univariate maximum likelihood and EM methods on data corresponding to the Paris-Lyon high-speed line provide the estimations $\hat{a}_1 = 0.0355, \hat{b}_1 = 19.19, \hat{a}_2 = 0.0387, \hat{b}_2 = 29.72, \hat{\rho} = 0.5262$. Usual thresholds are $s_1 = 0.9$ for NL and $s_2 = 0.75$ for NT. With these values, $\tau^{(1)}$, $\tau^{(2)}$ and τ^B are plotted in Figure 3 with respect of x_2 when x_1 is fixed ($x_1 = 0.4$). In that case $\tau^{(1)} = 150$.

This figure shows that taking into account the single information $x_1 = 0.4$ may lead to too late maintenance actions. As an example, if $x_2 = 0.4$, one has $\tau^B = 134.7$ (and $\tau^{(2)} = 152.9$). The preventive maintenance action based only on NL is consequently scheduled 15 days too lately. If $x_2 = 0.5$, one obtains $\tau^B = 95.9$ ($\tau^{(2)} = 97.5$) and the maintenance action is undertaken 54 days too late. If $x_2 = 0.6$, one obtains $\tau^B = 47.1$ ($\tau^{(2)} = 47.2$) and this is 103 days too late.

Concluding this section, one can finally observe that if x_1 is not too close to x_2, the value $\tau^U = \min\left(\tau^{(1)}, \tau^{(2)}\right)$ seams reasonable for maintenance scheduling (see Figure 3), contrary to the currently used $\tau^{(1)}$, which may entail large delay in its planning (more than 100 days in our example). If x_1 is close to x_2, the values of τ^U and τ^B have the same order of magnitude, however with $\tau^U > \tau^B$, so that the preventive maintenance action is again planned too lately (15 days in the example).

6 CONCLUSION

A bivariate Gamma process has been used to model the development of two deterioration indicators. Different estimation methods have been proposed for the parameters and tested on simulated data. Based on these tests, the

best estimators seem provided by univariate likelihood maximization for the marginal parameters and by an EM algorithm for the correlation coefficient.

Preventive maintenance scheduling has then been studied for a system that deteriorates according to a bivariate Gamma process. In particular, it has been shown that, given an observed bivariate deterioration level, the optimal time to maintenance is increasing with dependence. It has been proven that the optimal time to maintenance is always shorter when taking into account the dependence between the two deterioration indicators than when considering them separately (or only considering one of them).

Finally, a bivariate Gamma process has been used to study a real track maintenance problem. The application shows that when both observed deterioration indicators are close to each other, the bivariate process gives safer results for maintenance scheduling than both univariate processes considered separately or one single univariate process, with the same order of magnitude in each case however. When the observed deterioration indicators are clearly different, considering one single univariate process as it is done in current track maintenance, may lead to clearly unadaquate results. This application to real data of railway track deterioration hence shows the interest of a bivariate model for a correct definition of a maintenance strategy.

Bibliography

[1] C. Meier-Hirmer. *Modèles et techniques probabilistes pour l'optimisation des stratégies de maintenance. Application au domaine ferroviaire*. PhD thesis, Université de Marne-la-Vallée, 2007.

[2] J. M. van Noortwijk. A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1):2–21, 2009.

[3] M. Abdel-Hameed. A gamma wear process. *IEEE Transactions on Reliability*, 24(2):152–153, 1975.

[4] A. Grall, L. Dieulle, C. Bérenguer, and M. Roussignol. Continuous-time predictive-maintenance scheduling for a deteriorating system. *IEEE Transactions on Reliability*, 51(2):141–150, 2002.

[5] D. Zuckerman. Optimal replacement policiy for the case where the damage process is a one-sided Lévy process. *Stochastic Processes and their Applications*, 7:141–151, 1978.

[6] F. A. Buijs, J. W. Hall, J. M. van Noortwijk, and P. B. Sayers. Time-dependent reliability analysis of flood defences using gamma processes. In G. Augusti, G. I. Schuëller, and M. Ciampoli, editors, *Safety and Reliability of Engineering Systems and Structures; Proceedings of the Nineth International Conference on Structural Safety and Reliability (ICOSSAR), Rome, Italy, 19-23 June 2005*, pages 2209–2216, Rotterdam, 2005. Millpress.

[7] E. Çinlar, Z. P. Bažant, and E. Osman. Stochastic process for extrapolating concrete creep. *Journal of the Engineering Mechanics Division*, 103(EM6): 1069–1088, 1977.

[8] R. P. Nicolai, R. Dekker, and J. M. van Noortwijk. A comparison of models for measurable deterioration: an application to coatings on steel structures.

Reliability Engineering and System Safety, 92(12):1635–1650, 2007.

[9] C. Meier-Hirmer, G. Riboulet, F. Sourget, and M. Roussignol. Maintenance optimisation for system with a gamma deterioration process and intervention delay: application to track maintenance. *Journal of Risk and Reliability*, to appear, 2009.

[10] M. S. Joshi and A. M. Stacey. Intensity Gamma: a new approach to pricing portfolio credit derivatives. *Risk Magazine*, 6, 2006.

[11] F. Dufresne, H. U. Gerber, and E. S. W. Shiu. Risk theory with the gamma process. *ASTIN Bulletin*, 21(2):177–192, 1991.

[12] M. J. Newby and C. T. Barker. A bivariate process model for maintenance and inspection planning. *International Journal of Pressure Vessels and Piping*, 83(4):270–275, 2006.

[13] J. Kallsen and P. Tankov. Characterization of dependence of multidimensional Lévy processes using Lévy copulas. *Journal of Multivariate Analysis*, 97:1551–1572, 2006.

[14] L. Devroye. *Non-Uniform Random Variate Generation*. Springer, 2006.

[15] J. M. van Noortwijk and M. D. Pandey. A stochastic deterioration process for time-dependent reliability analysis. In M. A. Maes and L. Huyse, editors, *Proceedings of the Eleventh IFIP WG 7.5 Working Conference on Reliability and Optimization of Structural Systems, Banff, Canada, 2-5 November 2003*, pages 259–265, London, 2004. Taylor & Francis Group.

[16] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

An adaptive condition-based maintenance policy with environmental factors

Estelle Deloux and Bruno Castanier* – Ecole des Mines de Nantes, Nantes, France, Christophe Bérenguer – Université de Technologie de Troyes, Troyes, France

Abstract. This paper deals with the construction and optimisation of accurate condition-based maintenance policies for cumulative deteriorating systems. In this context, the system condition behavior can be influenced by different environmental factors which contribute to increasing or reducing the degradation rate. The observed condition can deviate from the expected condition if the degradation model does not embrace these environmental factors. Moreover, if more information is available on the environment variations, the maintenance decision framework should take advantage of this new information and update the decision. The question is how shall we model the decision framework for this? A gamma process-degradation model with randomized parameters is proposed to model the influence of the random environment on the system behavior. An adaptive maintenance policy is constructed which takes into account the environmental changes. The mathematical framework is presented here and a numerical experiment is conducted to highlight the benefit of our approach.

1 INTRODUCTION

Many manufacturing processes or structural systems suffer increasing wear with usage or age and are subject to random failures resulting from this deterioration and most of them are maintained or repairable systems. Appropriate maintenance actions such as inspection, local repair, and replacement should be done to protect manufacturing processes or structural systems from failure. However, the decisions depend on the quality of the model which represents the system subject to time-dependent degradation. The system modelling allows to have access to the "a priori"behavior of the system in terms of probability occurrence. Time-dependent degradation can be modelled in several ways [1]. In recent years, stochastic models with a rich probabilistic structure and simple methods for statistical inference

*corresponding author: IRCCyN, Ecole des Mines de Nantes, La chantrerie, 4 rue Alfred Kastler, BP 20722, 44307 Nantes, France; telephone: +33-(0)2 51858312, fax: +33-(0)2 51858349, e-mail: bruno.castanier@emn.fr

(e.g. gamma process models) have emerged due to modern developments in computational technology and the theory of stochastic processes.

The gamma process appears to be a very good candidate for degradation models because the sample path of a gamma process embraces both minute and large jumps [1, 2]. The degradation may develop in a very slow, invisible fashion due to daily usage and at some points in time it may grow very quickly when some traumatic event happen. Finally, the gamma processes is more versatile than a random variable model for stationary degradation processes [3], because it takes into account temporal uncertainty to better model when the variability in degradation is high. A recent overview of the gamma process in maintenance is given by Van Noortwijk [4].

Maintenance decisions regarding the time and frequency of inspection, repair and replacement are complicated by environmental uncertainty associated with the degradation of systems. Although many stochastic models of degradation with applications have been proposed [3], the impact of environmental uncertainty on maintenance optimisation has been lacking in the engineering literature. For a deteriorating system, an age-based maintenance policy is easier to implement than a condition-based maintenance policy (CBM), but CBM has proven its efficiency in terms of economic benefits and also in terms of system safety performance [5]. When environmental factors significantly impact the degradation of the system and this impact can be captured, it could be of interest to propose adaptive CBM policies as functions of the environmental variations [6, 7]. In their studies, [6, 7] assume definitive changes in the deterioration process parameters after a non-observable variation in the environment. Their model allows a shift to a new policy when the environment is supposed to have evolved. In [1], the changes are assumed to be reversible depend just on the stress level. The random evolution of the environment is modelled by a 2-state, continuous-time Markov chain and several policies allowing several shifts are proposed. Nevertheless, we underline the difficulty of obtaining the associated criterion and the applicability of such policies in the industrial context.

We develop in this work a new adaptive CBM framework for a dynamic deteriorating system, which allows only one potential shift in policies if the gap between expectation and observation is quite large. The direct observable environmental-stress process is modelled by a 2-state continuous-time Markov chain and the environmental impact on the system is modelled by deteriorating speed variations. The shift to a new policy will be done if the cumulative time elapsed in one environment becomes greater than an optimized threshold.

The remainder of this paper is as follows. In Section 2, the failure process and the relationship between deterioration level and stress covariate are presented. Section 3 is devoted to the construction of the new maintenance policy based on the system deterioration level and the stress variable to benefit such information. In Section 4, the numerical resolution of the cost criterion is briefly presented and a numerical example is proposed to

highlight the benefit of the new policy. Section 5 will discuss different extensions of this work.

2 DESCRIPTION OF THE FAILURE PROCESS

We consider a single-unit system subject to one failure mechanism evolving in a stressful environment. This section is devoted to describing the system failure process, the evolution of the stress and the relationship between these two processes. The system failure presented here is the same as the one presented in [1].

2.1 Stochastic deterioration model

The condition of the system at time t can be summarized by a scalar aging variable X_t [1, 8, 9, 10] whose variance increases as the system deteriorates. X_t can be the measure of a physical parameter linked to the resistance of a structure (e.g., length of a crack). The initial state corresponds to a perfect working state, $X_0 = 0$. The system fails when the aging variable is greater than a predetermined threshold L. The threshold L can be seen as a deterioration level which must not be exceeded for economical or security reasons. Let us model the degradation process $(X_t)_{(t\geq 0)}$ by a stationary gamma process where the increment of a degradation on a given time interval δt is gamma distributed. The associated probability density function is then:

$$f_{\alpha\delta t,\beta}(x) = \frac{1}{\Gamma(\alpha\delta t)}\beta^{\alpha\delta t}x^{\alpha\delta t-1}e^{-\beta x}I_{\{x\geq 0\}}(x), \tag{1}$$

where $I_A(x) = 1$ if $x \in A$ and 0 otherwise. We will not discuss here the several statistical properties of the gamma distribution nor the accuracy of the gamma process in the modelling of cumulative-deteriorating systems. We refer the interested reader to the survey of the application of gamma processes in maintenance [4] on the applicability of gamma processes in maintenance optimisation for many civil engineering structures.

2.2 Stress process

Let us assume that the system is subject to an environmental stress that can be external to the system (e.g., temperature, vibrations, ...) or a direct consequence of the system operating mode (e.g., internal vibrations, internal temperature, etc). We consider that the environmental condition at time t can be summarized with a single binary covariate $(Y_t)_{(t\geq 0)}$. We assume that Y_t is an indicator of the environmental evolution, i.e. it does not model the environment but only indicates if the system is in a stressed condition or not ($Y_t = 1$ if the system is stressed and 0 otherwise). The time intervals between successive state changes are exponentially distributed with parameter λ_0 (respectively, λ_1) for the transit to the non-stressed state from the stress state (respectively, stressed to non-stressed state). At each time $t \geq 0$, as the stressed state does not depend on the deterioration state, the probability of being in the stressed state in the steady-state is $1-p = \frac{\lambda_0}{\lambda_0+\lambda_1}$.

2.3 Impact of the stress process on the system deterioration

If the system evolves in a stressful environment, let us consider that the deterioration behavior can be impacted by this environment. We assume if $Y_t = y$ (with $y = 0$ or 1), $X_y(\delta t) \sim Ga(\alpha_0 e^{\gamma y}\delta t, \beta)$ [11, 12] where γ measures the influence of the covariate on the deterioration process.

Thus, we assume that the system is subject to an increase in the deterioration speed while it is under stress (i.e. while $Y_t = 1$), then the system deteriorates according to its nominal mode (while $Y_t = 0$). The parameters of the deterioration process when the system is non-stressed are $\alpha_0\delta t(\alpha = \alpha_0)$ and β and when the system is under stress $\alpha_1\delta t$ and β with $\alpha_1\delta t = \alpha_0 e^{\gamma y}\delta t$.

In average, the mean of the shape parameter $\bar{\alpha}$ is $\alpha_0(1+(1-p)(e^{\gamma}-1))$ and β can be estimated by using the maximum likelihood estimation. The parameter $\gamma > 0$ is an acceleration factor and can be obtained with the accelerated life testing method.

The model presented here is particularly adapted to an "internal" environment linked to observable missions profiles. For example, we can consider a motor which works according to its normal speed but it may be necessary to increase the production and thus to increase the speed of the motor. The time in the normal speed and in the accelerated speed is random, but it is measurable. This model can also be used for modelling road degradation which is based on the proliferation and growth of cracks. Moreover, environmental factors impact the road degradation, for example, extreme temperatures tends to increase it and it is possible to know the average time spent in extreme conditions.

Figure 1 sketches the different deterioration phases due to random evolution of the environment.

3 DEFINITION AND EVALUATION OF THE MAINTENANCE POLICY

In this section, two ways to integrate stress information in the decision framework are evaluated. The first is a static one in the sense that the decision rules are fixed. We will show that this policy, hereafter referred as *Policy 0*, mimics the classical inspection/replacement CBM policy with the stationary deterioration parameters $(\bar{\alpha}, \beta)$. The second policy is a dynamic one in the sense that the decision parameters can be updated according to the environmental condition. In section 3.3 we derive the long-run average maintenance cost per unit of time.

3.1 Structure of the static maintenance policy (Policy 0)

The cumulative deterioration level X_t is observed only through costly inspections. Let c_{ix} be the unitary inspection cost. Even if non-periodic inspection strategies are optimal [3], a periodic strategy is proposed here. The benefit of such an assumption is a reduced number of the decision

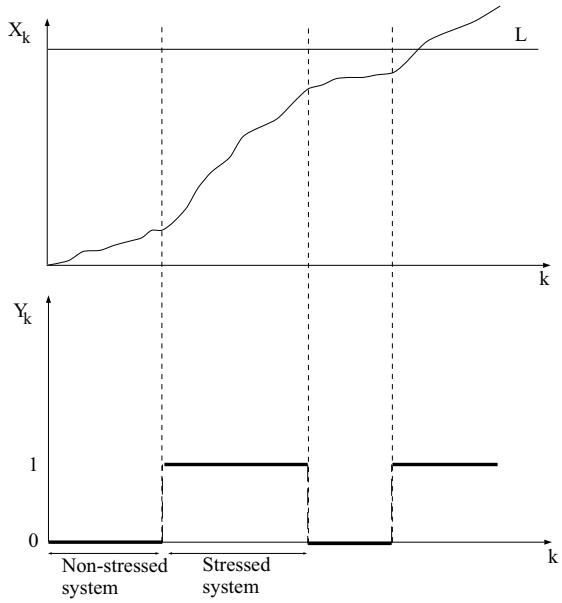


FIGURE 1. Evolution of the deterioration process impacted by the stress process

parameters, and an easier implementation of the approach in an industrial context. This inspection is assumed to be perfect in the sense that it reveals the exact deterioration level X_t.

During an inspection, a replacement can take place to renew the system if it is failed (corrective replacement) or to prevent the failure (preventive replacement). We assume the unitary cost of a corrective replacement c_c is composed of all the direct and indirect costs incurred by this maintenance action. Only the unitary unavailability cost c_u multiplied by the time the system is failed has to be added to c_c. The decision rule for a preventive replacement is the classical control limit rule: if ξ is the preventive replacement threshold, a preventive replacement is performed during the inspection on X_t if the deterioration level belongs to the interval (ξ, L). Let c_p be the preventive replacement cost ($c_p < c_c$).

Hence, the decision parameters which should be optimized in order to minimize the long-run maintenance cost are:

- The inspection period τ_0 which allows for balancing the cumulative inspection cost, earlier detection and prevention of a failure;
- The preventive maintenance threshold ξ which reduces cost by the prevention of a failure.

Finally, this approach corresponds to a classical maintenance policy taking

the stationary parameters $(\bar{\alpha} = \alpha_0(1 + (1-p)(e^{\gamma} - 1)), \beta)$.

This maintenance policy is denoted Policy 0 hereafter. An illustration of this maintenance policy is presented in Figure 2.

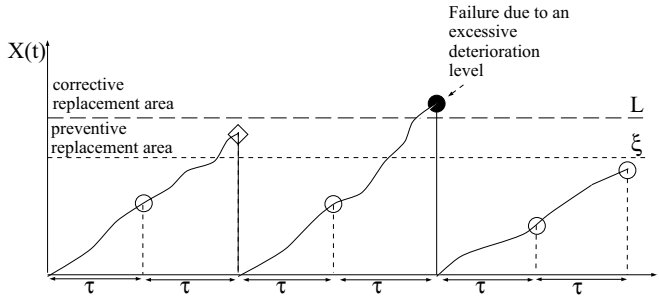


FIGURE 2. Evolution of the deterioration process and the stress process when the system is maintained

3.2 Structure of the dynamic maintenance policy (Policy 1)

Previously, for the Policy 0, only the information given by the deterioration level has been used, but it can be useful to adapt the decision for inspection and replacement with the observed time elapsed in the different operating conditions. In [1], we have investigated a panel of different CBM strategies taking into account the stress information. The main conclusions are that the updating of the decision according to the stress information should be continuously monitored by a new parameter as a function of the time proportion elapsed in the stressed state. Nevertheless, both optimisation and industrial implementation of such approaches are not easy.

In this section we develop a model free from limits of the models proposed in [1]. We propose a new maintenance policy (denoted Policy 1 hereafter) which still offers the opportunity to adapt the decision function to the time elapsed in the stress state. We still consider that the environment state is continuously monitored, but the number of updatings is limited: only one potential change is allowed in an inspection period. Both inspection interval and preventive replacement threshold can be updated. Before the description of the updating rule, let $r(t)$ be the actual time elapsed in the stressed state and $\bar{r}(t)$ the average of the time elapsed in the stressed state. $r(t)$ follows a k-Erlang law with parameter $\lambda_r = \frac{1}{1-p} = \frac{\lambda_0+\lambda_1}{\lambda_0}$ which leads to a discretisation of the time. Let us denote $\bar{r}_1(t)$ and $\bar{r}_2(t)$ two decision thresholds. The updating rule in the k^{th} inspection interval, $t_l \in (t_{k-1}, t_k)$, follows:

- while $r(t_l) \in (\bar{r}_1(t_l), \bar{r}_2(t_l))$, the decision rule is based on the policy 0, i.e (τ_0, ξ);

- if $r(t_l) < r_1(t_l)$, the (τ_0, ξ) rule is immediately and definitively replaced with (τ_1, ξ). Hence, the inspection will be differ from $t_k + \tau_0$ to $t_k + \tau_1$ and a preventive replacement will be performed if $X_{t_k+\tau_1} > \xi$.
- if $r(t_l) > r_2(t_l)$, the (τ_0, ξ) rule is immediately and definitively replaced with (τ_2, ξ).

After an inspection the next inspection planned is always planned τ_0 units of time later and is re-evaluated depending on $r(t)$.

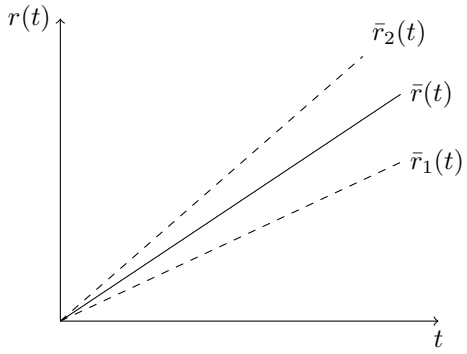


FIGURE 3. Evolution of the mean time elapsed in the stress state

3.3 Cost-based criterion for maintenance performance evaluation

In the case of policy 0, the maintenance decision parameters τ_0 and ξ should be optimized in order to minimize the long-run maintenance cost, but the cost criterion is obtained using the same reasoning as in the case of the dynamic maintenance policy. Thus, only the description of the cost evaluation in this last case is developed. For the dynamic maintenance policy, all the maintenance decision parameters are fixed $(\tau_0, \tau_1, \tau_2, \xi)$, optimized with the policy 0, and only the long-run maintenance cost needs to be estimated. By using classical renewal arguments, see e.g. [8], the maintenance criterion is expressed on a renewal cycle S defined by two consecutive replacements. Hence, we have:

$$\bar{C}_\infty(\tau_0, \tau_1, \tau_2, \xi) = \lim_{t\to\infty} \frac{C(t)}{t} = \frac{\mathbb{E}(C(S))}{\mathbb{E}(S)} \tag{2}$$

if $C(.)$ is the cumulative cost function, S is the first replacement date and $\bar{C}_\infty$ is the expected long-run maintenance cost. Nevertheless the calculation of these two expectations is not trivial in our case and we want to reduce the interval of calculation to the semi-renewal cycle $[0, T_1]$, i.e. to a shorter

interval (between two consecutive inspections). Therefore all the possible trajectories on $[0, T_1]$ are only deterioration trajectories (which are characterized by the Gamma law). The following result is proved for a semi-regenerative process for which the embedded Markov chain has a unique stationary probability distribution π:

$$\bar{C}_\infty(\tau_0, \tau_1, \tau_2, \xi) = \lim_{t\to\infty} \frac{C(t)}{t} = \frac{\mathbb{E}_\pi(C(T_1))}{\mathbb{E}_\pi(T_1)} \tag{3}$$

In return for the simplifications induced, this result requires to prove the existence of the stationary law of the Markov chain and to identify it. This study is not developed in this paper but the reasoning is the same as the one presented in [13].

The cost $C(T_1)$ is composed of the different inspections, replacements and unavailability costs and is written:

$$\bar{C}_\infty(\tau_0, \tau_1, \tau_2, \xi) = \frac{c_{ix}\mathbb{E}_\pi(N_{ix}(T_1)) + c_p + (c_c - c_p)\mathbb{P}_\pi(X_{T_1} > L) + c_u\mathbb{E}_\pi(D_u(T_1))}{\mathbb{E}_\pi(T_1)}, \tag{4}$$

where $N_{ix}(t)$ is the number of planned inspections before t and $D_u(t)$ the unavailability time before t.

4 NUMERICAL RESULTS

4.1 Evaluation of the stationary law

In order to evaluate the stationary law of the semi-regenerative process, we study the system evolution scenarios. We identify the possible trajectories of the process conditionally to the deterioration level at the beginning of a semi-renewal cycle (i.e. before the maintenance operation characterized by the value y) in order to reach the value x at the end of the cycle (i.e. before the maintenance operation). Between two consecutive inspections, the deterioration law of the system is only a function of the accumulated deterioration on this time interval. We identify six exclusive scenarios (in the case of the dynamic maintenance policy) allowing to pass of y in x:

- scenario 1: A preventive or corrective replacement is performed ($y \geq \xi$). After this maintenance, the system is new ($z = 0$) and the degradation process law is given by:
  - scenario 1.1: $f^{\tau_0}(x)\mathbb{P}(\forall t_l, t_l \in [0, \tau_1 - 1], \bar{r_1}(t_l) < r(t_l) < \bar{r_2}(t_l))$
  - scenario 1-2: $f^{\tau_1}(x)\mathbb{P}(\forall t_l, t_l \in [0, \tau_1 - 1], \bar{r_1}(t_l) \geq r(t_l))$
  - scenario 1-3: $f^{\tau_2}(x)\mathbb{P}(\forall t_l, t_l \in [0, \tau_1 - 1], r(t_l) \geq \bar{r_2}(t_l))$
- scenario 2: An inspection is performed without replacement ($y < \xi$). After this action, the system degradation is unchanged $z = y$ and the degradation process law is given by:

- scenario 2-1: $f^{\tau_0}(x-y)\mathbb{P}(\forall t_l, t_l \in [0, \tau_1 - 1], \bar{r_1}(t_l) < r(t_l) < \bar{r_2}(t_l))$
- scenario 2-2: $f^{\tau_1}(x-y)\mathbb{P}(\forall t_l, t_l \in [0, \tau_1 - 1], \bar{r_1}(t_l) \geq r(t_l))$
- scenario 2-3: $f^{\tau_2}(x-y)\mathbb{P}(\forall t_l, t_l \in [0, \tau_1 - 1], r(t_l) \geq \bar{r_2}(t_l))$

By using the total probability law, we obtain the stationary density function of the evolution process on a semi-renewal cycle:

$$\pi(x) = \int_{\xi}^{+\infty} \pi(y)dy \Big[f^{(\tau_0)}(x)(e^{-\lambda_r\lambda_0} - e^{-\lambda_r\lambda_1})^{\tau_1-1} + f^{(\tau_1)}(x)e^{-\lambda_r\lambda_1} \dots$$
$$\sum_{i=1}^{\tau_1-2}(e^{-\lambda_r\lambda_0} - e^{-\lambda_r\lambda_1})^i + f^{(\tau_2)}(x)(1 - e^{-\lambda_r\lambda_0}) \sum_{i=1}^{\tau_1-2}(e^{-\lambda_r\lambda_0} - e^{-\lambda_r\lambda_1})^i \Big] \dots$$
$$+ \int_0^{\xi} \pi(y) \Big[f^{(\tau_0)}(x-y)(e^{-\lambda_r\lambda_0} - e^{-\lambda_r\lambda_1})^{\tau_1-1} + f^{(\tau_1)}(x-y)e^{-\lambda_r\lambda_1} \dots$$
$$\sum_{i=1}^{\tau_1-2}(e^{-\lambda_r\lambda_0} - e^{-\lambda_r\lambda_1})^i + f^{(\tau_2)}(x-y)(1 - e^{-\lambda_r\lambda_0}) \sum_{i=1}^{\tau_1-2}(e^{-\lambda_r\lambda_0} - e^{-\lambda_r\lambda_1})^i \Big] dy \tag{5}$$

with

$$f^{(\tau_0)}(x) = \frac{1}{\Gamma(\bar{\alpha}(\tau_0 + r(\tau_0)e^{\gamma}))} \beta^{\bar{\alpha}(\tau_0 + r(\tau_0)e^{\gamma})} x^{\bar{\alpha}(\tau_0 + r(\tau_0)e^{\gamma})-1} e^{\beta x}$$

The stationary density function π (cf. Figure 4) is computed by numerical integration. Hence, the long-run expected maintenance cost per unit of time is numerically achievable.

4.2 Numerical example

This subsection is devoted to comparing the economic performance of the two proposed policies. We arbitrarily fix the maintenance data and the operations costs to the following values: the deterioration parameter $\alpha = 0.5$, $\beta = 20$, $\gamma = 2$; the stress parameter, $\bar{r}(t_l) = 0.666t_l$, $\bar{r}_1(t_l) = 0.6t_l$, $\bar{r}_2(t_l) = 0.714t_l$, the failure level $L = 2$, the maintenance costs $c_c = 100$, $c_p = 20$, $c_{ix} = 5$, $c_u = 50$. For the Policy 0, the optimized value of the inspection period is 18 when $\bar{r}(t_l) = 0.666t_l$ (respectively $\tau_1* = 20$ for $\bar{r}_1(t_l) = 0.6t_l$ and $\tau_2* = 16$ for $\bar{r}_2(t_l) = 0.714t_l$). The optimal cost obtained with the static policy, the Policy 0, is 2.023 and the one obtained with the Policy 1 is 1.915 which corresponds to a benefit of 5.333%. Policy 1 takes the advantage here to propose an adaptive inspection interval to the real proportion of time elapsed in the stress state. Even if this scheme is more complicated to implement than the static one it improves the economical performance.

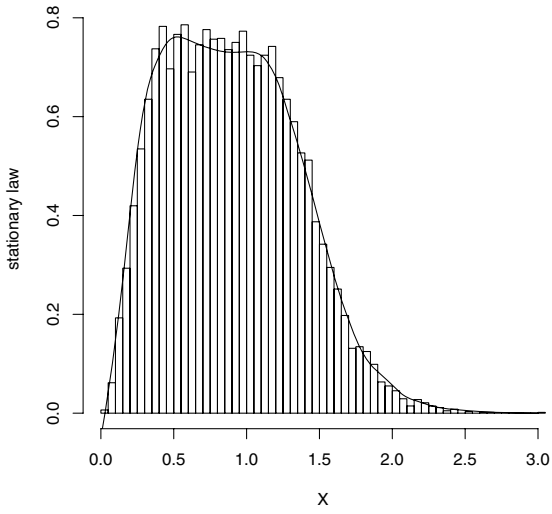


FIGURE 4. Stationary density function (histogram: the stationary density function π obtained by simulations, curve: the stationary density function π obtained numerically)

5 DISCUSSION

The main interest of this work is the construction and the evaluation of maintenance policies for continuously deteriorating systems subject to environmental influence. Taking into account the environment makes the system deterioration behavior dynamic. This is a common assumption in an industrial context. The relationship between the system performance and the associated operating environment has been modelled as an accelerator factor for deterioration and as a binary variable. A cost criterion has been numerically evaluated to highlight the performance of the different maintenance strategies and the benefits to consider the opportunity to adapt the current decision according to the history of the system.

Even if the last proposed structure for maintenance decision framework has shown interesting performance, a lot of research remains to be done. A sensitivity analysis when maintenance data varies should be performed. Moreover, for the moment we fix $\bar{r}_1(t)$ and $\bar{r}_2(t)$ but it could be interesting to optimise them in order to improve the economic benefits. Furthermore, in practice it is exceptional that the system deterioration level can be measured directly and, very often, only information correlated at the deterioration level is observable. It is thus necessary to develop conditional maintenance policies for which the decision is taken from this imperfect,

partial information. Several research tracks are possible. For example, hidden Markov processes can be adapted to model the indirectly observable deterioration. It is also possible to consider that we observe a process correlated in the deterioration process. In our case, we could take the stress, and reconstruct the real state of the system from the observations before making a decision of maintenance. Additionally, due to the model assumptions in this paper, we propose a system for which the result if the system is in the stressed state followed by a non-stressed state produces in average the same degradation as the opposite (if the time elapsed in the stressed state and in the non-stressed state are preserved). But for many systems this reciprocity is not true, this assumption should be relaxed. Furthermore, it could be interesting to transfer the lessons of the case known environment with uncertainty to the case "non observable" environment and to compare the estimation of the time elapsed in the stressed and non-stressed state to the expectations.

Acknowledgments

This study is part of the SBaDFoRM (State-Based Decision For Road Maintenance) project financed by the region Pays de la Loire (France). The authors gratefully acknowledge the helpful comments from the reviewers.

Bibliography

[1] E. Deloux, B. Castanier, and C. Bérenguer. Comparison of health monitoring strategies for a gradually deteriorating system in a stressfull environment. In T. Kao, E. Zio, and V. Ho, editors, *International Conference on Probabilistic Safety Assement and Management (PSAM), Hong Kong, China, May 2008*. PSAM, 2008.

[2] N. Singpurwalla. Gamma processes and their generalizations: an overview. In *Engineering Probabilistic Design and Maintenance for Flood Protection*, pages 67–73. Kluwer Academic, Dordrecht, 1997.

[3] X. Yuan. *Stochastic model of deterioation in nuclear power plant components*. PhD thesis, University of Waterloo, Ontario, Canada, 2007.

[4] J. M. van Noortwijk. A survey of the application of gamma processes in maintenance. *Reliability Engineering and System Safety*, 94(1):2–21, 2009.

[5] I. Gertsbakh. *Reliability Theory With Applications to Preventive Maintenance*. Springer, Berlin, 2000.

[6] M. Fouladirad, A. Grall, and L. Dieulle. On the use of on-line detection for maintenance of gradually deteriorating systems. *Reliability Engineering and System Safety*, 93:1814–1820, 2008.

[7] B. Saassouh, L. Dieulle, and A. Grall. Online maintenance policy for a deteriorating system with random change of mode. *Reliability Engineering and System Safety*, 92:1677–1685, 2007.

[8] S. Asmussen. *Applied Probability and Queues, Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons, Chichester, 1987.

[9] B. Castanier E. Deloux and C. Bérenguer. Combining statistical process control and condition-based maintenance for gradually deteriorating systems

subject to stress. In T. Aven and J. E. Vinnem, editors, *Risk, Reliability and Societal Safety, Proceedings of ESREL 2007 - European Safety and Reliability Conference 2007, Stavanger, Norway, 25-27 June 2007*, pages 265–272, London, 2007. Taylor & Francis.

[10] M. Rausand and A. Hoyland. *System Reliability Theory-Models, Statistical Methods, and Applications*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2004.

[11] B. Castanier, C. Bérenguer, and A. Grall. A sequential condition-based repair/replacement policy with non-periodic inspections for a system subject to continuous wear. *Applied Stochastic Models in Business and Industry*, 19 (4):327–347, 2003.

[12] Irving W. Burr. *Statistical quality control methods*. Marcel Dekker, New York, 1976.

[13] L. Dieulle, C. Bérenguer, A. Grall, and M. Roussignol. Sequential condition-based maintenance scheduling for a deteriorating system. *European Journal of Operational Research*, 150:451–461, 2003.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

Derivation of a finite time expected cost model for a condition-based maintenance program

Mahesh D. Pandey* and Tianjin Cheng – University of Waterloo, Waterloo, Canada

Abstract. The gamma process is a stochastic cumulative process that can be used to model a time-variant uncertain process. Professor van Noortwijk's research work played a key role in modeling degradation by gamma process and making it popular in engineering community. The maintenance optimization models mostly use the renewal theorem to evaluate the asymptotic expected cost rate and optimize the maintenance policy. However, many engineering projects have relative short and finite time horizon in which the application of the asymptotic formula becomes questionable. This paper presents a finite time model for computing the expected maintenance cost and investigates the suitability of the asymptotic cost rate formula.

1 INTRODUCTION

This paper considers the optimization of a condition-based maintenance (CBM) of components that are subjected to gradual degradation, such as material corrosion or creep. The theory of stochastic processes has provided a valuable framework to model temporal uncertainty associated with degradation. Since degradation in typical engineering components tends to be monotonic and cumulative over time, cumulative stochastic processes have been used to model the damage and predict reliability. The gamma process is an example of a stochastic cumulative process with a simple mathematical structure that provides an effective tool to model time-variant degradation.

Although the basic mathematical framework of the gamma process was developed in early seventies, Professor van Noortwijk should be given credit for introducing this model to civil engineering community [1, 2, 3]. A comprehensive review of the gamma process model and its applications was recently published by van Noortwijk [4]. Gamma process has been applied to model various types of degradation processes, such as creep in concrete [5], recession of coastal cliffs [6], deterioration of coating on steel structures [7], structural degradation [8] and wall thinning corrosion of pipes in nuclear power plants [9].

*corresponding author: Department of Civil and Environmental Engineering, University of Waterloo; 200 University Ave. West; Waterloo, ON, Canada N2L 3G1; telephone: +1-519 888 4567 35858, fax: +1-519 888 4349, e-mail: mdpandey@uwaterloo.ca.

The CBM policy considered on this paper involves periodic inspections to quantify the amount of degradation, an indicator of the condition, at different points in time. The component fails when degradation exceeds a critical value, d_F. Therefore it is desirable to make a preventive replacement (PR) as the degradation reaches a limit, d_P, which is less than d_F. The cost of replacement after failure (FR) is significantly higher than that associated with PR due to lack of spares, long outage and sudden disruption of services. So the objective of maintenance program is to determine the inspection interval, T, and damage level for PR, d_P, that would minimize the life cycle cost of operative this component.

Several variations of the CBM policy have been discussed in the literature, depending on whether or not the inspection schedule is periodic, inspection tools are perfect, failure detection is immediate, or repair duration is finite. Park [10] studied periodic CBM policy of a component subjected to stochastic continuous degradation. Park's model was extended in [11] by considering a random preventive replacement level of the damage. These two models assumed that failure is self-announced, i.e., an inspection is not needed to detect the failure. The case in which failure could only be detected through inspection was analyzed in [12]. Grall et al. [13] studied the case in which inspection is non-periodic and the preventive level is fixed. The case of imperfect inspection was analyzed by Kallen and van Noortwijk [14]. Castanier et al. [15] studied a type of maintenance policy in which both the future operation (replacement or imperfect repair) and the inspection schedule depend on the current degradation.

In most of the literature, the criterion for optimizing CBM is based on minimizing the expected cost per unit time, i.e., the *cost rate*. The computation of the cost rate is difficult as it involves computation of convolutions of different probability distributions. The renewal theorem provides a simple alternative to compute long term or asymptotic value of the expected cost rate [16, 17]. The asymptotic rate is the expected cost in one renewal cycle divided by the expected duration of the renewal cycle. However, many engineering projects have relatively short and finite time horizon in which the applicability of asymptotic formula becomes questionable.

This paper presents a finite time model for evaluating the expected cost associated with a periodic CBM policy. The solution approach is based on formulating the expected cost as a generalized renewal equation and the computations are done on high performance computers. The paper presents a case study involving CBM of piping systems in a nuclear plant. It is illustrated that the asymptotic formula over predicts the life cycle cost as compared to that obtained from the proposed finite time model.

This paper is organized as follows. Section 2 briefly describes the stationary gamma process model. Section 3 formulates the periodic CBM policy. The derivation of the proposed finite time cost model is presented in Section 4. An illustrative example is given in Section 5 and conclusions are summarized in Section 6.

2 GAMMA PROCESS DEGRADATION MODEL

Let $X(\tau)$ denote the degradation at time τ after the renewal of the component. $X(0)=0$ and $X(\tau)$ increases with τ. The component fails when $X(\tau) \geq d_F$. We use compact notations to denote the following probability terms: $P\{X(\tau) \leq x\} = P(\tau, x)$ and $P\{X(\tau_1) \leq x_1, X(\tau_2) \leq x_2\} = P(\tau_1, x_1; \tau_2, x_2)$, $\tau_1 \leq \tau_2$.

The degradation process, $X(\tau)$, is modeled as a continuous stationary gamma process, which is defined as follows. Recall that the probability density function of a gamma distributed random variable, Z, is given by:

$$g(z|\alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} z^{\alpha-1} e^{-z/\beta} \tag{1}$$

where α and β are the shape and scale parameters, respectively, and the complete gamma function is denoted as $\Gamma(u) = \int_0^{\infty} x^{u-1}e^{-x}dx$. The cumulative distribution function (CDF) of Z is [4]

$$G(z|\alpha, \beta) = P\{Z \leq z\} = \frac{\Gamma_{z/\beta}(\alpha)}{\Gamma(\alpha)}, \tag{2}$$

where $\Gamma_v(u) = \int_0^v x^{u-1}e^{-x}dx$ is an incomplete gamma function. The gamma process, $X(\tau), \tau \geq 0$, has the following properties [4]:

1. $X(0)=0$ with probability one;

2. Increments over an interval $\Delta\tau$ are gamma distributed with scale $\alpha\Delta\tau$ and shape β, i.e., $\Delta X(\Delta\tau) \equiv X(\tau+\Delta\tau) - X(\tau) \sim \text{Ga}(\alpha\Delta\tau, \beta)$; and

3. $X(\tau)$ has independent increments.

In case of gamma process, the following probability terms are introduced.

$$P(\tau, x) = G(x \,|\, \alpha\tau, \beta), \tag{3}$$

$$P(\tau_1, x_1; \tau_2, x_2) = \int_0^{x_1} G\big(x_2 - y \,\big|\, \alpha(\tau_2 - \tau_1), \beta\big) g(y \,|\, \alpha\tau_1, \beta) dy. \tag{4}$$

The distribution of the component lifetime, A, can be obtained as

$$F_A(a) = P\{A \leq a\} = P\{X(a) \geq d_F\} = 1 - \frac{\Gamma_{d_F/\beta}(\alpha a)}{\Gamma(\alpha a)} \tag{5}$$

Given degradation inspection data, the shape and the scale parameters of the gamma process can be estimated from the methods of maximum likelihood, moments and the Bayesian statistics [18, 7, 4].

3 CONDITION-BASED MAINTENANCE (CBM)

A component is inspected periodically at a constant time interval of T and the amount of degradation $X(\tau)$ is measured. The component fails when $X(\tau) > d_F$. The failure is immediately noticed and the component is replaced promptly. The component can be preventively replaced if $X(\tau)$ exceeds a pre-selected level of d_P at the time of inspection (see Figure 1a). Note that PR and FR denote preventive replacement and that upon failure, respectively.

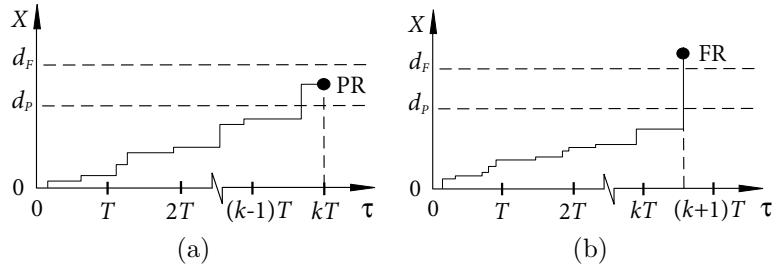


FIGURE 1. Types of replacement: (a) preventive replacement and (b) failure replacement

Let L be the length of the operation period and J be the type of replacement, $J \in \{\mathrm{PR}, \mathrm{FR}\}$. In case of $J = \mathrm{PR}$, L is only a multiple of the inspection interval T, i.e., T, $2T$, $\cdots$, kT, because PR only occurs at the time of inspection. As shown in Figure 1a, the probability of PR at an inspection time kT for any integer k can be evaluated as:

$$\begin{aligned} P\{L=kT\,,\, J=\mathrm{PR}\} &= P\big\{X\big((k-1)T\big) \le d_P\,,\, d_P < X(kT) \le d_F\big\} \\ &= P\big((k-1)T, d_P\,;\, kT, d_F\big) - P\big(kT, d_P\big). \end{aligned} \tag{6}$$

Since the failure can take place at any time in between the inspection intervals, the probability of failure replacement (FR) within an interval can be evaluated as (see Figure 1b):

$$\begin{aligned} P\{kT < L \le kT+h\,,\, J=\mathrm{FR}\} &= P\big\{X(kT) \le d_P\,,\, X(kT+h) > d_F\big\} \\ &= P\big(kT, d_P\big) - P\big(kT, d_P; kT+h, d_F\big). \end{aligned} \tag{7}$$

with $0 < h \le T$.

Denote the probability of PR at any time of inspection as

$$q_{\mathrm{PR},k} = P\{L=kT\,,\, J=\mathrm{PR}\}, \tag{8}$$

and the probability density function (PDF) of L when $J=\mathrm{FR}$ at $\tau = kT+h$ as

$$q_{\mathrm{FR}}(\tau) = \frac{dP\big\{kT < L \le kT+h\,,\, J=\mathrm{FR}\big\}}{dh}. \tag{9}$$

The next step is to derive expected length of the renewal cycle, L. No replacement before $(kT+h)$ means $X(kT) \leq d_P$ and $X(kT+h) \leq d_F$. Hence

$$P\{L > kT+h\} = P\big(kT, d_P\,;\, kT+h, d_F\big). \tag{10}$$

Then the PDF of L at $\tau = kT+h$ can be given as

$$q(\tau) = -\frac{dP\{L > kT+h\}}{dh} \tag{11}$$

and the expected value of L

$$\mathbf{E}\{L\} = \sum_{k=0}^{\infty} \int_0^T P\{L > kT+h\} dh. \tag{12}$$

The total life cycle cost includes costs of periodic inspections, preventive and failure replacements. Denote the unit cost of PR, FR and inspection by c_{PR}, c_{FR}, and c_{IN}, respectively. Note that $c_{\mathrm{PR}} \ll c_{\mathrm{FR}}$ is a common sense assumption in any CBM policy. An operating cycle, L, ends with a PR or FR, and the cost associated with any general cycle is denoted as $c(L, J)$. For any integer k and h, $0 < h \leq T$, total costs associated with cycles ending with PR or FR are given as

$$c(kT, \mathrm{PR}) = c_{\mathrm{PR}} + c_{\mathrm{IN}} k, \qquad c(kT+h\,,\,\mathrm{FR}) = c_{\mathrm{FR}} + c_{\mathrm{IN}} k. \tag{13}$$

The expected cost in one renewal cycle is computed as

$$\mathbf{E}\{c\} = \sum_{k=1}^{\infty} (c_{\mathrm{PR}} + c_{\mathrm{IN}} k)\, q_{\mathrm{PR},k} + \sum_{k=0}^{\infty} \int_{kT}^{(k+1)T} (c_{\mathrm{FR}} + c_{\mathrm{IN}} k)\, q_{\mathrm{FR}}(\tau) d\tau. \tag{14}$$

Note that Equations (12) and (14) can also be found in [10].

4 EVALUATION OF THE EXPECTED LIFE-CYCLE COST

Under the CBM policy, a series of pairs $\{L_i, J_i\}$, $i = 1, 2, \cdots$, cover the planning horizon of the CBM policy. It is assumed that the lifetime of all replacements are *iid* random variables and the time spent for replacement is negligible. Let S_n be the chronological time of occurrence of n^{th} replacement and $N(t)$ be the number of replacements up to time t. Then

$$S_n = \sum_{i=1}^{n} L_i, \quad N(t) = \max_{S_n \leq t} n.$$

$N(t)$ is a renewal process [16, 17] with renewal times $\{S_n\}$, $n = 1, 2, \cdots$. Denoting the cost of an ith renewal cycle as $c(L_i, J_i)$, the total cost up to t can be written as

$$C(t) = \sum_{i=1}^{N(t)} c(L_i, J_i) + c_{\mathrm{IN}} \left\lfloor \frac{t - S_{N(t)}}{T} \right\rfloor, \tag{15}$$

where $\lfloor * \rfloor$ means the floor function. The last term in the right hand side is the additional inspection cost in the interval $\big(S_{N(t)}\,,\,t\big]$. Given the first pair $\{L_1, J_1\} = \{l, j\}$, if $l < t$, the conditional expected cost can be formulated as

$$\begin{aligned}\mathbf{E}\Big\{C(t)\,\Big|\,L_1 = l, J_1 = j\Big\} &= \mathbf{E}\Big\{\text{Cost in } (0,l] + \text{Cost in } (l,t]\,\Big|\,L_1 = l, J_1 = j\Big\} \\ &= c(l,j) + \mathbf{E}\big\{C(t-l)\big\}. \end{aligned} \tag{16}$$

If $l > t$, only inspection cost incurs in $(0, t]$, such that

$$\mathbf{E}\left\{C(t)\,\middle|\,L_1 = l\right\} = c_{\text{IN}}\,\lfloor t/T \rfloor\,. \tag{17}$$

Conditioned on three mutually exclusive cases: $\{L_1 \le t, J_1 = \text{PR}\}$, $\{L_1 \le t, J_1 = \text{FR}\}$, and $\{L_1 > t\}$, $\mathbf{E}\{C(t)\}$ can be partitioned as follows

$$\begin{aligned}\mathbf{E}\big\{C(t)\big\} = &\sum_{k=1}^{\lfloor t/T \rfloor} \mathbf{E}\left\{C(t)\,\middle|\,L_1 = kT, J_1 = PR\right\} q_{\text{PR},\,k} + \cdots \\ &\sum_{k=0}^{\lfloor t/T \rfloor} \int_{\Delta_k} \mathbf{E}\big\{C(t)\,\big|\,L_1 = \tau, J_1 = FR\big\} q_{\text{FR}}(\tau) d\tau + \cdots \\ &\int_t^{\infty} \mathbf{E}\left\{C(t)\,\middle|\,L_1 = \tau\right\} q(\tau) d\tau \end{aligned} \tag{18}$$

where

$$\Delta_k = \begin{cases} \big(kT, (k+1)T\big], & \text{for } 0 \le k < \lfloor t/T \rfloor, \\ \big(\,\lfloor t/T \rfloor\, T, t\big], & \text{for } k = \lfloor t/T \rfloor, \end{cases} \quad \text{and} \quad \bigcup_{k=0}^{\lfloor t/T \rfloor} \Delta_k = (0, t].$$

Substituting Equations (13), (16) and (17) into Equation (18) gives

$$\begin{aligned}\mathbf{E}\big\{C(t)\big\} = &\sum_{k=1}^{\lfloor t/T \rfloor} \Big[\,(c_{\text{PR}} + c_{\text{IN}} k) + \mathbf{E}\big\{C(t - kT)\big\}\Big] q_{\text{PR},\,k} + \cdots \\ &\sum_{k=0}^{\lfloor t/T \rfloor} \int_{\Delta_k} \Big[\big(c_{\text{FR}} + c_{\text{IN}} k\big) + \mathbf{E}\big\{C(t-\tau)\big\}\Big] q_{\text{FR}}(\tau) d\tau + c_{\text{IN}} \left\lfloor \frac{t}{T} \right\rfloor \int_t^{\infty} q(\tau) d\tau \end{aligned} \tag{19}$$

Denoting $\mathbf{E}\{C(t)\}$ by $U(t)$, Equation (19) can be simplified as

$$U(t) = G(t) + \left[\sum_{k=0}^{\lfloor t/T \rfloor} U(t - kT) q_{\text{PR},\,k} + \int_0^t U(t-\tau) q_{\text{FR}}(\tau) d\tau \right], \tag{20}$$

where

$$\begin{aligned} G(t) = c_{\text{PR}} &\sum_{k=1}^{\lfloor t/T \rfloor} q_{\text{PR},k} + c_{\text{FR}} \int_0^t q_{\text{FR}}(\tau) d\tau + \cdots \\ &c_{\text{IN}} \left\{ \sum_{k=1}^{\lfloor t/T \rfloor} k \left[q_{\text{PR},k} + \int_{\Delta_k} q_{\text{FR}}(\tau) d\tau \right] + \lfloor t/T \rfloor \int_t^{\infty} q(\tau) d\tau \right\}. \end{aligned}$$

Note that the PDF of L when $J = \mathrm{PR}$ can be written as

$$q_{\mathrm{PR}}(\tau) = \sum_{k=1}^{\infty} q_{\mathrm{PR},k}\delta(\tau - kT),$$

$\delta(*)$ being the Dirac delta function, and $q(\tau) = q_{\mathrm{PR}}(\tau) + q_{\mathrm{FR}}(\tau)$. Equation (20) can be rewritten in a more compact form as

$$U(t) = G(t) + \int_0^t U(t-\tau)q(\tau)d\tau, \tag{21}$$

Equation (21) is a generalized renewal equation, which can be solved for $U(t)$ with the initial condition $U(0) = 0$. To compute $U(t)$, the time horizon is discretized in small intervals as v, $2v$, $\cdots$, and denoting $t = nv$, introduce discrete variables $U_i = U(iv)$, $G_i = G(iv)$, and $q_i = q(iv)$, $i = 1, 2, \cdots, n$. Equation (21) can be re-written in a discrete form as

$$U_0 = 0, \quad U_n = G_n + \sum_{i=1}^{n} U_{n-i}q_i, \quad \text{for } n \geq 1, \tag{22}$$

from which U_1, U_2, $\cdots$, can be computed in a recursive manner.

In case of an infinite time horizon ($t \to \infty$), the expected asymptotic cost rate converges to the ratio of the expected cost in one renewal interval to the expected length of the renewal cycle, i.e.,

$$u_\infty = \frac{\mathbf{E}\{c\}}{\mathbf{E}\{L\}}. \tag{23}$$

$\mathbf{E}\{c\}$ and $\mathbf{E}\{L\}$ can be obtained from equations (12) and (14), respectively. The expected cost over a time horizon t is then estimated as $C_\infty \approx t \times u_\infty$.

5 EXAMPLE

Flow accelerated corrosion (FAC) degradation is common in the heat transport piping system (PHTS) of nuclear power plants. The corrosion can be modelled as a stochastic gamma process [9]. The objective is to evaluate the life cycle cost associated with a condition-based maintenance of the piping system. The following information is gathered from the inspection data [19]. The initial wall thickness of the pipe is 6.50 mm. The minimum required wall thickness is 2.41 mm. The degradation level corresponding to failure is thus $d_F = 3.09$ mm. Using the inspection data regarding wall thickness measurements in a sample, the parameters of the gamma process were estimated as $\alpha = 1.13$/year and $\beta = 0.0882$ mm. The PDF of the lifetime obtained from Equation (5) is plotted in Figure 2. The mean lifetime is 31.63 years and the standard deviation 5.24 years.

Cost data are specified in a relative scale as $c_{\mathrm{IN}} = 0.1$, $c_{\mathrm{PR}} = 1$, and $c_{\mathrm{FR}} = 10$. The preventive replacement level is chosen as $d_P = 2.0$ mm based

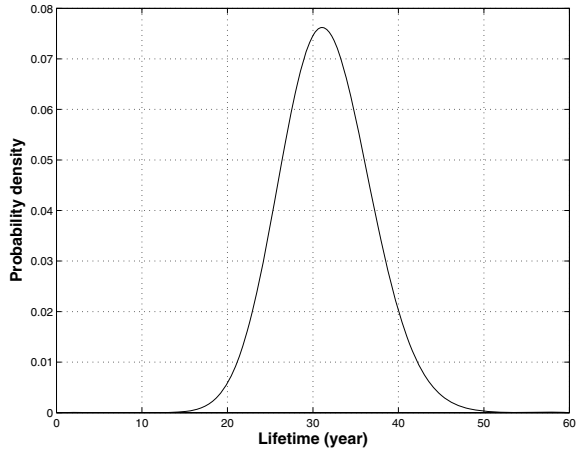


FIGURE 2. PDF of the lifetime

on a regulatory requirement and the planning horizon is taken as $t = 50$ years.

The expected cost from finite time model is computed using equation (21), and the asymptotic cost is a product of the asymptotic rate u_∞ with the length of the interval. u_∞ was computed from Equation (23). The variation of the total expected cost with respect to the inspection interval is plotted in Figure 3. The finite time model results in the optimal inspection interval of 9 years and corresponding minimum life cycle cost of 2.55 units. The asymptotic formula results in an optimal inspection interval of 7 years and the associated cost is 3.09 units, which is about 20% higher than that calculated from the finite time formula. This is the difference in the cost predicted by finite time and asymptotic formulas for one pipe section in the plant. Given that the Canadian reactor design consists of 380 to 480 pipe sections, this cost differential for the entire reactor would be quite large. This underscores the need for using the proposed finite time cost computation model for safety critical infrastructure systems.

The mathematical formulation is quite versatile and it can be used to optimize other parameters of the CBM plan. For example, the the damage level corresponding to the preventive replacement level, d_P, can be optimized for a given inspection interval.

6 CONCLUSIONS

This paper presents the derivation of the expected life-cycle cost associated with a periodic CBM policy in a finite time horizon. The degradation in a component is modeled as stochastic gamma process. The derivation is

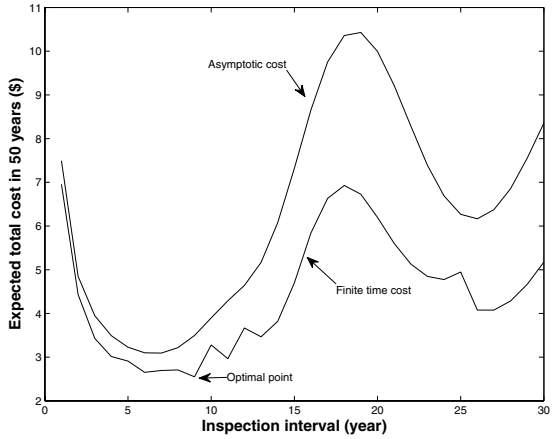


FIGURE 3. Expected cost versus inspection interval over a 50 year period

based on formulating a generalized renewal equation for the expected cost and computing convolutions using high performance computers.

The paper highlights the fact that the asymptotic expected cost can be a rather crude approximation of the real cost in a finite time horizon. The paper presents a case study involving CBM of piping systems in a nuclear plant, which illustrates that the asymptotic formula over predicts the life cycle cost by 20% as compared to that obtained from the proposed finite time model. Given that a plan contains a large fleet of piping components, the over prediction by asymptotic formula can be substantial, which paints a pessimistic picture of the life cycle cost at the plant level. It is concluded that the finite time model should be used for a realistic evaluation and optimization of the CBM policy for safety critical infrastructure systems. The formulation presented in the paper can be extended to other types of maintenance policies.

Acknowledgments

We acknowledge financial support for this study provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University Network of Excellence in Nuclear Engineering (UNENE) through an Industrial Research Chair program on Risk-Based Life Cycle Management of Engineering Systems at the University of Waterloo.

Bibliography

[1] J. M. van Noortwijk and P. H. A. J. M. van Gelder. Optimal maintenance decisions for berm breakwaters. *Structural Safety*, 18(4):293–309, 1996.

[2] J. M. van Noortwijk and H. E. Klatter. Optimal inspection decisions for the block mats of the Eastern-Scheldt barrier. *Reliability Engineering and System Safety*, 65:203–211, 1999.

[3] J. M. van Noortwijk, J. A. M. van der Weide, M. J. Kallen, and M. D. Pandey. Gamma process and peaks-over-threshold distributions for time-dependent reliability. *Reliability Engineering and System Safety*, 92:1651–1658, 2007.

[4] J. M. van Noortwijk. A survey of the application of gamma processes in maintenance. *Reliability Engineering and System Safety*, 94:2–21, 2009.

[5] E. Çinlar, Z. P. Bažant, and E. Osman. Stochastic process for extrapolating concrete creep. *Journal of the Engineering Mechanics Division*, 103(6):1069–1088, 1977.

[6] J. W. Hall, I. C. Meadowcroft, E. M. Lee, and P. H. A. J. M. van Gelder. Stochastic simulation of episodic soft coastal cliff recession. *Coastal Engineering*, 46(3):159–174, 2002.

[7] R. P. Nicolai, R. Dekker, and J. M. van Noortwijk. A comparison of models for measurable deterioration: An application to coatings on steel structures. *Reliability Engineering and System Safety*, 92(12):1635–1650, 2007.

[8] D. M. Frangopol, M. J. Kallen, and J. M. van Noortwijk. Probabilistic models for life-cycle performance of deteriorating structures: review and future directions. *Progress in Structural Engineering and Materials*, 6(4):197–212, 2004.

[9] X. X. Yuan, M. D. Pandey, and G. A. Bickel. A probabilistic model of wall thinning in CANDU feeders due to flow-accelerated corrosion. *Nuclear Engineering and Design*, 238(1):16–24, 2008.

[10] P. S. Park. Optimal continuous-wear limit replacement under periodic inspections. *IEEE Transactions on Reliability*, 37(1):97–102, 1988.

[11] M. B. Kong and K. S. Park. Optimal replacement of an item subject to cumulative damage under periodic inspections. *Microelectronics and Reliability*, 37(3):467–472, 1997.

[12] M. Abdel-Hameed. Inspection and maintenance policies for devices subjected to deterioration. *Advances in Applied Probability*, 19(4):917–931, 1987.

[13] A. Grall, C. Bérenguer, and L. Dieulle. A condition-based maintenance policy for stochastically deteriorating systems. *Reliability Engineering and System Safety*, 76:167–180, 2002.

[14] M. J. Kallen and J. M. van Noortwijk. Optimal maintenance decisions under imperfect inspection. *Reliability Engineering and System Safety*, 90(2-3): 177–185, 2005.

[15] B. Castanier, C. Bérenguer, and A. Grall. A sequential condition-based repair/replacement policy with non-periodic inspections for a system subjected to continuous wear. *Applied Stochastic Models in Business and Industry*, 19 (4):327–347, 2003.

[16] W. L. Smith. Regenerative stochastic processes. *Proceedings of the Royal Society*, 232:6–31, 1955.

[17] D. R. Cox. *Renewal Theory*. Methuen, London, 1962.

[18] I. V. Basawa and P. J. Brockwell. A note on estimation for gamma and stable processes. *Biometrica*, 67(1):234–236, 1980.

[19] M. D. Pandey, D. Lu, and D. Komljenovic. The impact of probabilistic models in life cycle analysis of pipes in nuclear plants. Accepted for publication in ASME Journal of Engineering for Gas Turbines and Power, 2009.

Risk and Decision Analysis in Maintenance Optimization and Flood Management
M.J. Kallen and S.P. Kuniewski (Eds.)
IOS Press
© 2009. The authors and IOS Press. All rights reserved.

A discussion about historical developments in stochastic modeling of wear

Hans van der Weide* – Delft University of Technology, Delft, the Netherlands and Mahesh D. Pandey – University of Waterloo, Waterloo, Canada

Abstract. In this paper we study models for cumulative damage of a component caused by shocks occurring randomly in time, following a historical approach. The damage caused by a shock, is also of random nature. A very well-known model is the compound renewal process, with the compound Poisson process as a special case. These models play an important role in maintenance analysis and cost calculations. In these models the times at which shocks occur and the damage caused by the shock are assumed to be independent. But very often this is not realistic, the damage will depend on the time since the last shock, in some engineering applications it is even a deterministic function of the time since the last shock. Also, the results are often asymptotic. We will develop a model which allows dependence between damage and time since the last shock. We will calculate Laplace transforms of the interesting quantities and show how these can be inverted to get probability distributions for finite time horizons.

1 INTRODUCTION

In this paper we study models for cumulative damage of a component, more in particular models that are time-dependent. The component fails if the cumulative damage exceeds some critical threshold. The threshold represents the resistance of the component, which is degrading in time. If the degradation cannot be neglected and if we want to insert uncertainty, the threshold has to be modelled as a stochastic process as well.

The first models of this type are studied in Mercer and Smith's paper [1] that was published in 1959. In this paper the authors introduce a stochastic model for the wear of a conveyor belt. In the terminology of Mercer and Smith, the damage is modelled by a one-dimensional random walk process in which the steps are positive and occur randomly at mean rate m and such that the sizes of the steps are independent with probability density f on $[0, \infty)$. Basically, the authors derive a formula for the probability density of

*corresponding author: Department of Applied Mathematics, Delft University of Technology, P.O. Box 5301, NL-2600 GA Delft, The Netherlands; telephone: +31-(0)15 27 87286, e-mail: j.a.m.vanderweide@tudelft.nl

the time needed to reach a fixed barrier, as well as the asymptotic moments of this time as the height of the barrier goes to infinity.

In a paper [2] dating from 1965, Morey gives a generalization of the model of Mercer and Smith. He proposes to replace the Poisson process, that was used to model the times at which damages occur, with a renewal process and he calls his model a compound renewal process. Compound renewal processes have been introduced by Smith in a survey paper [3] published in 1958. Also, Morey proposes to use nonparametric models for the jump size, such as distributions with monotone failure rate. In the paper, bounds are derived for the mean and the variance of the first time the total damage reaches a random barrier X. The second part of the paper deals with the special case that the total damage is modelled as a compound Poisson process in which jump-size has a Pólya-frequency density of order 2. In this case it is shown that the hitting time of the barrier has an increasing failure rate and a monotone likelihood ratio.

We will study the probability distribution and the moments of the first time τ_h that the total damage exceeds some given threshold h. In Section 2 we discuss Mercer and Smith's model. The explicit formula for all moments of τ_h is new. Next, we characterize the probability distribution of τ_h via a double Laplace transform, that can be transformed back numerically to get the probability distribution. We also discuss the case where the wear is not only caused by shocks, but also by a nearly continuous abrasion. In Section 3 we propose a model which is a generalization of a model introduced by Morey [2], where we allow dependencies between the jumps and the inter-occurrence times. The theoretical results about this model are not new. They have been derived in more general form in the mathematical literature, see the work of Shanthikumar and Sumita [4] and [5]. Our contributions are the asymptotic properties for the moments of τ_h and the numerical inversion of the formula for its double Laplace transform that can be used for numerical inversion. We also show the importance of the assumption about the dependence between time and damage by giving an example based on the bivariate exponential distribution, see [6], Chapter 5.

2 MERCER AND SMITH'S MODEL

In their paper [1], Mercer and Smith present a model for the wear of conveyor belting. In modern terminology, the total damage is modelled by a compound Poisson process $\{X(t) : t \geq 0\}$ with intensity $\lambda = m$ and random jump size with probability density f. So the times at which shocks (belting damages) occur are modelled by a homogeneous Poisson process $N = \{N(t) : t \geq 0\}$, or stated otherwise, the times between consecutive shocks are stochastically independent and exponentially distributed with mean $1/m$. Damage is only caused by the shocks. The severity of the damage is stochastic and is modelled by an i.i.d. sequence of nonnegative random variables (Y_k), independent of the damage times process N. So,

assuming that damage is additive and that the system does not recover, the total damage at time t is given by

$$X(t) = \sum_{k=1}^{N(t)} Y_k.$$

The belt is considered to be worn out completely if the total damage reaches a certain level. Define, for a given, non-random damage level $h > 0$, the first time that X exceeds this level by

$$\tau_h = \min\{t \geq 0 : X(t) > h\},$$

and let η_k denote the total damage from the first k shocks

$$\eta_k = Y_1 + \cdots + Y_k, \quad k = 1, 2, \ldots$$

As usual we define $\eta_0 \equiv 0$. Since the process X has right continuous, increasing sample paths,

$$\tau_h \leq t \quad \Longleftrightarrow \quad X(t) > h, \tag{1}$$

and it follows from independence of the process N and the sequence $(Y_k)_{k\geq 1}$ that

$$\mathbb{P}(\tau_h > t) = \mathbb{P}(X(t) \leq h) = e^{-mt} \sum_{k=0}^{\infty} \frac{(mt)^k}{k!} p_k(h),$$

where, for $k \geq 0$, $p_k(h) = \mathbb{P}(\eta_k \leq h)$. For practical purposes, the infinite sum can be truncated. A rough upper bound for the error, if we approximate $\mathbb{P}(\tau_h > t)$ with the sum of the first n terms, is given by $(mtp_1(h))^{n+1}/(n+1)!$.

It follows from the expression for $\mathbb{P}(\tau_h > t)$ that τ_h is a continuous random variable with probability density function:

$$\begin{aligned} g_h(t) &= me^{-mt} \sum_{k=0}^{\infty} \frac{(mt)^k}{k!} (p_k(h) - p_{k+1}(h)) \\ &= me^{-mt} \sum_{k=0}^{\infty} \frac{(mt)^k}{k!} \mathbb{P}(\eta_k \leq h < \eta_{k+1}). \end{aligned} \tag{2}$$

The rth moment of τ_h (possibly infinite) is given by

$$\mathbb{E}(\tau_h^r) = r \int_0^{\infty} t^{r-1} \mathbb{P}(\tau_h > t)\, dt = \frac{r}{m^r} \sum_{k=0}^{\infty} \frac{(r+k-1)!}{k!} p_k(h).$$

To derive an alternative expression for the rth moment of τ_h, let $\tilde{N}$ be the renewal process associated to the sequence $(Y_k)_{k\geq 1}$. Then

$$p_k(h) = \mathbb{P}(Y_1 + \ldots + Y_k \leq h) = \mathbb{P}(\tilde{N}_h \geq k),$$

and it follows that

$$\begin{aligned}
\mathbb{E}(\tau_h^r) &= \frac{r}{m^r}\sum_{k=0}^{\infty}\frac{(r+k-1)!}{k!}\mathbb{P}(\tilde{N}_h \geq k) \\
&= \frac{r}{m^r}\sum_{k=0}^{\infty}\frac{(r+k-1)!}{k!}\sum_{i=k}^{\infty}\mathbb{P}(\tilde{N}_h = i) \\
&= \frac{r}{m^r}\sum_{i=0}^{\infty}\left(\sum_{k=0}^{i}\frac{(r+k-1)!}{k!}\right)\mathbb{P}(\tilde{N}_h = i) \\
&= \frac{1}{m^r}\sum_{i=0}^{\infty}(i+1)\cdots(i+r)\mathbb{P}(\tilde{N}_h = i) \\
&= \frac{1}{m^r}\mathbb{E}\left((\tilde{N}_h+1)\cdots(\tilde{N}_h+r)\right).
\end{aligned}$$

So we have the following well-known elegant result for the moments of τ_h:

$$\mathbb{E}(\tau_h^r) = \frac{1}{m^r}\mathbb{E}\left((\tilde{N}_h+1)\cdots(\tilde{N}_h+r)\right), \tag{3}$$

see Hameed and Proschan [7] or Marshall and Shaked [8]. Applying the Key Renewal Theorem to the renewal process $\tilde{N}$, see Chapter 8 in Tijms [9], we find asymptotic expansions of the first two moments of τ_h as $h \to \infty$. Let $\mu_k = \mathbb{E}(Y_1^k)$. If $\mu_2 < \infty$, then

$$\lim_{h\to\infty}\left(\mathbb{E}(\tau_h) - \frac{h}{m\mu_1}\right) = \frac{\mu_2}{2m\mu_1^2}. \tag{4}$$

If $\mu_3 < \infty$, then

$$\lim_{h\to\infty}\left(\mathbb{E}(\tau_h^2) - \left\{\frac{1}{m^2\mu_1^2}h^2 + \frac{2\mu_2}{m^2\mu_1^3}h\right\}\right) = \frac{9\mu_2^2 - 4\mu_1\mu_3}{6m^2\mu_1^4}. \tag{5}$$

Since the sample paths of the cumulative damage process are right continuous, we have $X(\tau_h) > h$. Define the overshoot by $\gamma_h = X(\tau_h) - h$. Note that

$$\gamma_h = \sum_{i=0}^{\tilde{N}(h)+1} Y_i - h,$$

so the overshoot γ_h is the excess or residual life at time h of the renewal process $\tilde{N}$. It follows that

$$\mathbb{E}(\gamma_h) = \mu_1(1 + M_1(h)) - h,$$

where $M_1(t) = \mathbb{E}(\tilde{N}(t))$ is the renewal function associated to $\tilde{N}$ and

$$\lim_{h\to\infty}\mathbb{E}(\gamma_h) = \frac{\mu_2}{2\mu_1},$$

see Tijms [9]. For the second moment we have the formula

$$\mathbb{E}(\gamma_h^2) = \mu_1^2 M_2(t) + (\mu_2 + \mu_1^2 - 2\mu_1 t)M_1(t) + t^2 - 2\mu_1 t,$$

where $M_2(t) = \mathbb{E}(\tilde{N}^2(t))$. Also the asymptotic expansions of the second moment and the distribution function are well-known:

$$\lim_{h\to\infty} \mathbb{E}(\gamma_h^2) = \frac{\mu_3}{3\mu_1},$$

and

$$\lim_{h\to\infty} \mathbb{P}(\gamma_h \le x) = \frac{1}{\mu_1}\int_0^x (1 - F(y))\, dy, \quad x \ge 0.$$

An alternative way to characterize the probability distribution of τ_h is via its Laplace transform.

$$\begin{aligned}
\mathbb{E}\left(e^{-u\tau_h}\right) &= \int_0^\infty e^{-ut} g_h(t)\, dt \\
&= \sum_{k=0}^\infty \int_0^\infty m e^{-(u+m)t} \frac{(mt)^k}{k!} \mathbb{P}(\eta_k \le h < \eta_{k+1})\, dt \\
&= \sum_{k=0}^\infty \left(\frac{m}{u+m}\right)^{k+1} \mathbb{P}(\eta_k \le h < \eta_{k+1}).
\end{aligned}$$

This expression for the Laplace transform of τ_h is still not attractive to find the probability distribution of τ_h, even not numerically. Since double (or two-dimensional) Laplace transforms can be treated numerically without problems, we take the Laplace transform with respect to the variable h as well:

$$\begin{aligned}
&\int_0^\infty e^{-sh} \mathbb{E}\left(e^{-u\tau_h}\right)\, dh \\
&\qquad = \sum_{k=0}^\infty \left(\frac{m}{u+m}\right)^{k+1} \int_0^\infty \mathbb{P}(\eta_k \le h < \eta_{k+1}) e^{-sh}\, dh. \quad (6)
\end{aligned}$$

Here it is useful to use that $\mathbb{P}(\eta_k \le h < \eta_{k+1}) = p_k(h) - p_{k+1}(h)$. Now

$$\int_0^\infty p_k(h) e^{-sh}\, dh = \int_0^\infty \int_0^h g_k(x)\, dx e^{-sh}\, dh = \frac{1}{s}\{\mathcal{L}_f(s)\}^k,$$

where g_k is the density of $\eta_k = Y_1 + \ldots + Y_k$ and $\mathcal{L}_f(s) = \mathbb{E}\left(e^{-sY_1}\right)$ the Laplace transform of the jump height. It follows that

$$\int_0^\infty \mathbb{P}(\eta_k \le h < \eta_{k+1}) e^{-sh}\, dh = \frac{1}{s}\{\mathcal{L}_f(s)\}^k \left(1 - \mathcal{L}_f(s)\right),$$

hence the double Laplace transform of τ_h is given by

$$\int_0^\infty e^{-sh}\mathbb{E}\left(e^{-u\tau_h}\right)\,dh = \frac{m\left(1-\mathcal{L}_f(s)\right)}{s\left\{u+m\left(1-\mathcal{L}_f(s)\right)\right\}}. \tag{7}$$

So the double Laplace transform is determined by the intensity m of the Poisson process of the times at which the damages occur and the Laplace transform of the severity of the damage. We continue with three examples.

2.1 Example

Let the jump-size be constant, $Y \equiv d$. The renewal process $\tilde{N}$ associated with the jumps is then deterministic: $\tilde{N}(t) = k$ if $kd \leq t < (k+1)d$. So $\tilde{N}(t) = \lfloor t/d \rfloor$. It follows that

$$g_h(t) = me^{-mt}\sum_{k=0}^{\infty}\frac{(mt)^k}{k!}\mathbb{P}(\tilde{N}(h)=k) = me^{-mt}\frac{(mt)^n}{n!},$$

where $n = \lfloor h/d \rfloor$. So the distribution of τ_h is a gamma distribution and

$$\mathbb{E}(\tau_h^r) = \frac{1}{m^r}\frac{(r+n)!}{n!}, \quad n = \lfloor h/d \rfloor.$$

2.2 Example

Let the jump-size distribution be exponential with parameter λ. Without loss of generality we may assume that $\lambda = 1$. Then

$$p_k(h) = \mathbb{P}(Y_1 + \ldots + Y_k \leq h) = \int_0^h e^{-s}\frac{s^{k-1}}{(k-1)!}\,ds, \quad k \geq 1.$$

It follows by partial integration that

$$p_{k+1}(h) = -e^{-h}\frac{h^k}{k!} + p_k(h), \quad k \geq 0,$$

so, the probability density of τ_h is given by

$$\begin{aligned} g_h(t) &= me^{-mt}\sum_{k=0}^{\infty}\frac{(mt)^k}{k!}(p_k(h) - p_{k+1}(h)) \\ &= me^{-mt-h}\sum_{k=0}^{\infty}\frac{(hmt)^k}{(k!)^2} \\ &= me^{-mt-h}I_0(2\sqrt{hmt}), \end{aligned}$$

where I_0 denotes the modified Bessel function of the first kind with series expansion

$$I_0(z) = \sum_{k=0}^{\infty}\frac{(\frac{1}{4}z^2)^k}{(k!)^2}.$$

The renewal process $\tilde{N}$ associated with the jump sizes is in this case a homogeneous Poisson with intensity 1. It follows from formula (3) that

$$\mathbb{E}(\tau_h) = \frac{h+1}{m} \quad \text{and} \quad \mathrm{Var}(\tau_h) = \frac{2h+1}{m^2}.$$

Note that in this example $\mu_k = k!$ and it follows that

$$\mathbb{E}(\tau_h) = \frac{h}{m\mu_1} + \frac{\mu_2}{2m\mu_1^2}$$

and

$$\mathbb{E}(\tau_h^2) = \frac{1}{m^2\mu_1^2}h^2 + \frac{2\mu_2}{m^2\mu_1^3}h + \frac{9\mu_2^2 - 4\mu_1\mu_3}{6m^2\mu_1^4}.$$

The first two moments of the overshoot $\gamma_h = X(\tau_h) - h$ are equal to

$$\mathbb{E}(\gamma_h) = h + 1 \quad \text{and} \quad \mathbb{E}(\gamma_h^2) = h^2 + 2h.$$

2.3 Example

Let the jump-size distribution be a gamma distribution $\Gamma(\beta, 1)$, i. e.

$$f(y) = e^{-y}\frac{y^{\beta-1}}{\Gamma(\beta)}, \quad y \geq 0 \quad \text{and} \quad \mathcal{L}_f(s) = \left(\frac{1}{1+s}\right)^{\beta}.$$

It follows that

$$\int_0^\infty \mathbb{E}\left(e^{-u\tau_h}\right) e^{-sh}\, dh = \frac{m\left((1+s)^\beta - 1\right)}{s\left\{u(1+s)^\beta + m\left((1+s)^\beta - 1\right)\right\}}.$$

Unfortunately, it is not possible to get a nice analytical expression for the inverse of this double Laplace transform. We use methods from [10] for numerical inversion of the double Laplace transform. In Figure 1 the probability density of τ_1 is displayed for $m = 1$ and several values of β. For $\beta = 1$, the jump-size distribution is exponential, see the last Example.

We conclude this Section with a discussion of the case of a moving boundary. In their paper [1], Mercer and Smith discuss the case where the barrier h is replaced by the moving barrier $h - \lambda t$, where $\lambda > 0$ is a constant. The term λt can be considered as the wear caused by nearly continuous abrasion. It follows that the first time that this level is exceeded is now given by

$$\tau_{h,\lambda} = \min\{t \geq 0 : X(t) > h - \lambda t\}.$$

The analysis of the distribution of $\tau_{h,\lambda}$ in [1] is based on an approximation of the wear caused by abrasion by adding to $X(t)$ an independent Poisson process with intensity m_1 and jump-size distribution concentrated in x_1. Noting that the limit is the process λt if $m_1 \to \infty$, $x_1 \to 0$ such that $m_1 x_1 \to \lambda$, they use the result for Poisson processes with fixed barriers. Instead of

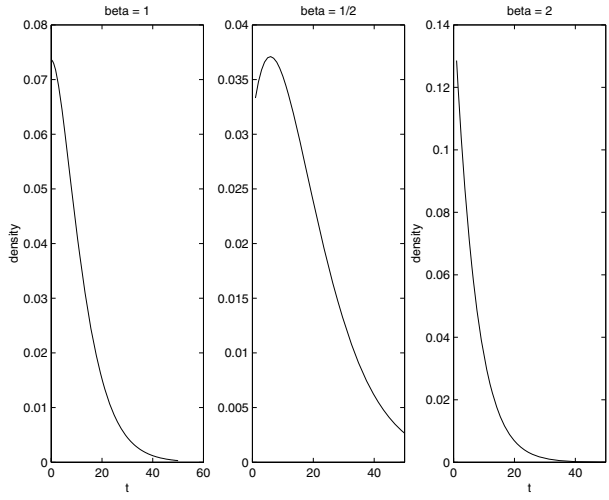


FIGURE 1. Probability densities of τ_1 for $m = 1$ and $\beta = 1$, $\beta = 1/2$ and $\beta = 2$ respectively.

this approach, we can calculate the two-dimensional Laplace transform of $\tau_{h,\lambda}$. It turns out that

$$\mathbb{P}(\tau_{h,\lambda} > t) = e^{-mt} \sum_{k=0}^{\infty} \frac{(mt)^k}{k!} p_k(h - \lambda t),$$

and

$$\int_0^{\infty} e^{-sh} \mathbb{E}\left(e^{-u\tau_{h,\lambda}}\right) \, dh = \frac{\lambda s + m(1 - \mathcal{L}_f(s))}{s(u + \lambda s + m(1 - \mathcal{L}_f(s)))}.$$

This formula can be inverted numerically which will give us all the information about the wear process that we need.

3 GENERALIZED MOREY MODEL

We discuss now a more general model, which is a generalization of a model introduced by Richard C. Morey. In Mercer's paper [1], the shocks occur according to a homogeneous Poisson process. Morey proposes to use a renewal process to describe the occurrence of the shocks. So the wear is modelled as a so-called compound renewal process. Compound Poisson processes have been introduced by W. L. Smith in [3]. Compound renewal processes are extensively used as models in economical and actuarial applications.

Let $0 = S_0 < S_1 < S_2 < \ldots$ be the times at which shocks occur. We will model these times as a renewal process. This means that the times T_i between successive shocks, i.e.

$$T_i = S_i - S_{i-1}, \quad i = 1, 2, \ldots,$$

are independent and identically distributed, strictly positive random variables. The cumulative distribution function of the inter-occurrence times T_j will be denoted by F. If $F(x) = 1 - e^{-mx}$ is the cumulative distribution function of the exponential distribution, we have the case discussed in [1]. For all $t \geq 0$, we denote by $N(t)$ the number of shocks during the time interval $[0, t]$, so

$$N(t) = \max\{j \mid S_j \leq t\}.$$

Let the damage occurring at time S_j be given by the random variable Y_j. We will assume that the sequence $\{(T, Y), (T_j, Y_j), j \geq 1\}$ of random vectors is an i.i.d. sequence and we will denote the cumulative distribution function of the random vector (T, Y) by H :

$$H(x, y) = \mathbb{P}(T \leq x,\, Y \leq y).$$

So the severity of the damage and the time since the last shock may be dependent. Note that $F(x) = H(x, +\infty)$. Assuming additivity of damage and no recovery, the total damage $X(t)$ occurred during the time interval $[0, t]$ can now be expressed by the formula,

$$X(t) = \begin{cases} \sum_{j=1}^{N(t)} Y_j & \text{if } T_1 \leq t, \\ 0 & \text{if } T_1 > t. \end{cases} \tag{8}$$

The process $\{X(t), t \geq 0\}$ is in the literature also known as a renewal reward process, see [9].

Denote, as before, by τ_h the time at which the process X crosses for the first time the level $h > 0$. Here it is in general not possible to give a useful formula for the probability distribution of τ_h, so we try to calculate the Laplace transform of τ_h. By partial integration and formula (1) we get

$$\mathbb{E}(e^{-u\tau_h}) = 1 - \int_0^\infty u e^{-ut} \mathbb{P}(X(t) \leq h)\, dt. \tag{9}$$

To do anything with this formula, we need to know the probability $\mathbb{P}(X(t) \leq h)$, which is the same as $\mathbb{P}(\tau_h > t)$, the probability that we are trying to find. Since in the case of a compound Poisson process it turned out to be useful to consider the double Laplace transform, we will use here the same approach. By partial integration,

$$\int_0^\infty e^{-sh} \mathbb{P}(X(t) \leq h)\, dh = \frac{1}{s} \mathbb{E}\left(e^{-sX(t)}\right), \tag{10}$$

Using (9) and (10), we get for the double Laplace transform

$$\begin{aligned} \int_0^\infty e^{-sh} \mathbb{E}(e^{-u\tau_h})\, dh &= \int_0^\infty e^{-sh} \left(1 - \int_0^\infty u e^{-ut} \mathbb{P}(X(t) \leq h)\, dt\right) dh \\ &= \frac{1}{s} \left\{1 - \int_0^\infty u e^{-ut} \mathbb{E}\left(e^{-sX(t)}\right) dt\right\}. \end{aligned} \tag{11}$$

Denote the Laplace-Stieltjes transforms of F and H by $\mathcal{L}_F$ and $\mathcal{L}_H$ respectively, i.e

$$\mathcal{L}_F(u) = \int_0^\infty e^{-ut}\, dF(t) = \mathbb{E}\left(e^{-uT}\right)$$

and

$$\mathcal{L}_H(u,s) = \int_0^\infty \int_0^\infty e^{-ux-sy}\, dH(x,y) = \mathbb{E}\left(e^{-uT-sY}\right),$$

for all $u, s \geq 0$. Then we have the following formula for the double Laplace transform of τ_h.

Theorem 3.1 *Let $\{X(t), t \geq 0\}$ be a renewal reward process. Then, for $u, s > 0$,*

$$\int_0^\infty e^{-sh}\mathbb{E}(e^{-u\tau_h})\, dh = \frac{\mathcal{L}_F(u) - \mathcal{L}_H(u,s)}{s(1-\mathcal{L}_H(u,s))}. \tag{12}$$

Proof. We calculate the righthand side of formula (11). Conditioning on the event $\{T_1 = x, C_1 = y\}$ we get,

$$\begin{aligned}
\mathbb{E}[e^{-sX(t)}] &= \int_0^\infty \int_0^t \mathbb{E}\left(e^{-sX(t)} \mid T_1 = x, C_1 = y\right) dH(x,y) + \cdots \\
&\quad + \int_0^\infty \int_t^\infty \mathbb{E}\left(e^{-sX(t)} \mid T_1 = x, C_1 = y\right) dH(x,y) \\
&= \int_0^\infty \int_0^t \mathbb{E}\left(e^{-s(y+X(t-x))}\right) dH(x,y) + (1 - F(t)). \qquad (13)
\end{aligned}$$

Multiplying the first term in the righthand side of formula (13) with ue^{-ut} and integrating with respect to t, we get

$$\begin{aligned}
&\int_0^\infty ue^{-ut}\left(\int_0^\infty \int_0^t \mathbb{E}\left(e^{-s(y+X(t-x))}\right) dH(x,y)\right) dt \\
&= \int_0^\infty \int_0^\infty e^{-sy}\left(\int_x^\infty ue^{-ut}\mathbb{E}\left(e^{-sX(t-x)}\right) dt\right) dH(x,y) \\
&= \int_0^\infty \int_0^\infty e^{-ux-sy}\left(\int_0^\infty ue^{-ur}\mathbb{E}\left(e^{-sX(r)}\right) dr\right) dH(x,y) \\
&= \mathcal{L}_H(u,s)\int_0^\infty ue^{-ut}\mathbb{E}\left(e^{-sX(t)}\right) dt. \qquad (14)
\end{aligned}$$

Multiplying the second term in the righthand side of formula (13) with ue^{-ut} and integrating with respect to t, we get

$$\int_0^\infty ue^{-ut}(1 - F(t))\, dt = 1 - \mathcal{L}_F(u). \tag{15}$$

So, multiplying the lefthand side of equation (13) with ue^{-ut} and integrating with respect to t, and substituting the formulas (14) and (15) in the

righthand side, we get

$$\int_0^\infty ue^{-ut}\mathbb{E}\left(e^{-sX(t)}\right)\,dt = \mathcal{L}_H(u,s)\int_0^\infty ue^{-ut}\mathbb{E}\left(e^{-sX(t)}\right)\,dt + 1 - \mathcal{L}_F(u),$$

which implies that

$$\int_0^\infty ue^{-ut}\mathbb{E}\left(e^{-sX(t)}\right)\,dt = \frac{1-\mathcal{L}_F(u)}{1-\mathcal{L}_H(u,s)}.$$

Substitution of this formula in equation (11) gives the result. □

Theorem 3.1 has been proven in more general form in Theorem 2.A1 in Sumita and Shanthikumar [5].

As special cases consider where T and Y are independent with distributions F and G respectively. This is the model studied in Morey's paper [2]. The model is in this case known as a compound renewal process, see Smith [3].

$$\int_0^\infty e^{-sh}\mathbb{E}(e^{-u\tau_h})\,dh = \frac{\mathcal{L}_F(u)(1-\mathcal{L}_G(s))}{s(1-\mathcal{L}_F(u)\mathcal{L}_G(s))}. \tag{16}$$

If T and Y are independent with distributions $F(x) = 1 - e^{-mx}$ and G respectively. Then

$$\mathcal{L}_F(u) = \frac{m}{m+u}$$

and

$$\int_0^\infty e^{-sh}\mathbb{E}(e^{-u\tau_h})\,dh = \frac{m\,(1-\mathcal{L}_G(s))}{s\,\{u+m\,(1-\mathcal{L}_G(s))\}}, \tag{17}$$

which is in agreement with the earlier derived formula (7).

Our results can also be applied in calculations for discounted life-cycle costs. Here the random variables Y_k represent the cost (notation C_k) of the kth repair. An important special case is the case where the cost $C = c(T)$ is given as a (non-random) function of the time since the last repair. This is the case that T and C are totally dependent. Here, the double integral in the calculation of $\mathcal{L}_H(u,s)$ reduces to a single integral:

$$\mathcal{L}_H(u,s) = \int_0^\infty e^{-ux-sc(x)}\,dF(x).$$

3.1 Example

As an example of the application of Theorem 3.1, consider the case where the shocks occur according to a homogeneous Poisson process with intensity m and with exponentially distributed damage Y. We will use Marshall & Olkin's bivariate distribution to introduce a dependence structure between the damage and the time since the last shock, see [11] and [6]. This bivariate exponential distribution can be described as follows.

$$T = \min(U,V), \quad Y = \min(W,V),$$

where U, V, W are independent random variables with exponential distributions with parameters λ, μ and ν respectively. It follows that the joint survival probability of the pair (T, Y) is

$$\bar{F}(x, y) = \mathbb{P}(T > x, Y > y) = e^{-(\lambda x - \nu y - \mu \max(x,y))}$$

with exponential marginal distributions

$$\mathbb{P}(T > x) = e^{-(\lambda+\mu)x}, \quad \mathbb{P}(Y > y) = e^{-(\mu+\nu)y}.$$

This distribution is characterized among the bivariate distributions with exponential marginal distributions by the following bivariate version of the memoryless property:

$$\mathbb{P}(T > x + z, Y > y + z \mid T > z, Y > z) = \mathbb{P}(T > x, Y > y)$$

for all $x \geq 0$, $y \geq 0$, $z \geq 0$. See [6], Chapter 5, where more information can be found.

The Laplace transform of the random vector (T, Y) is given by

$$\begin{aligned}
\mathcal{L}_H(u, s) &= \mathbb{E}\left(e^{-uT - sY}\right) \\
&= \mathbb{E}\left(\int_0^\infty 1_{[T,\infty)}(x) u e^{-ux}\, dx \int_0^\infty 1_{[Y,\infty)}(y) s e^{-sy}\, dy\right) \\
&= \int_0^\infty \int_0^\infty u s e^{-ux-sy} \mathbb{P}(T \leq x, Y \leq y)\, dxdy \\
&= 1 - \frac{u}{\lambda + \mu + u} - \frac{s}{\mu + \nu + s} \\
&\qquad + \frac{us(\lambda + 2\mu + \nu + u + s)}{(\lambda + \mu + u)(\nu + \mu + s)(\lambda + \mu + \nu + u + s)}.
\end{aligned}$$

Substitution of this formula in (12) together with $\mathcal{L}_F(u) = m/(m+u)$ yields a formula for the double Laplace transform for the first hitting time of level h. Using the results from [10] we can invert this double Laplace transform to get the probability density of the first hitting time of a given level h. For the dependent case it seems that the density $f(t, h)$, given by

$$f(t, h)\, dt = \mathbb{P}(\tau_h \in dt)$$

is not differentiable at the point $t = h$, which causes problems for the numerical Laplace inversion. Here we use the window-function $1 - e^{-(x-h)}$ to smoothen this function.

Figure 2 contains the result of the calculation of the probability density of τ_1 for the case $\lambda = \mu = \nu = 1$. The histogram is obtained from 10^5 simulations of the process.

To compare this result with the probability density that we get if we assume independence of the shock and the time since the last shock (i.e. $\mu = 0$) we display, for $\lambda = \nu = 1$, the probability density and a simulation of τ_1 in Figure 3.

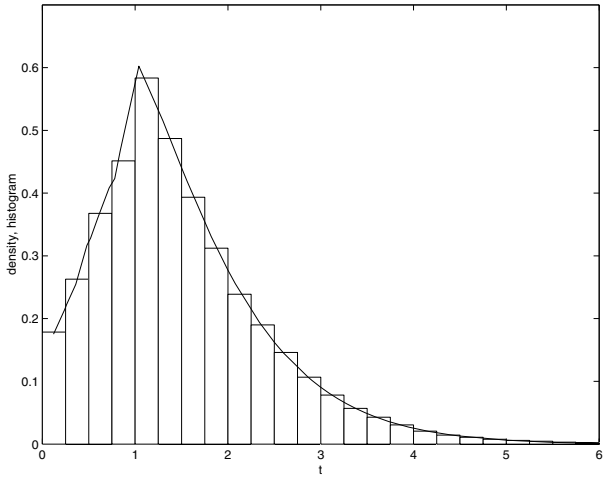


FIGURE 2. Density (numerical Laplace inversion) and Histogram (simulation) for the dependent case.

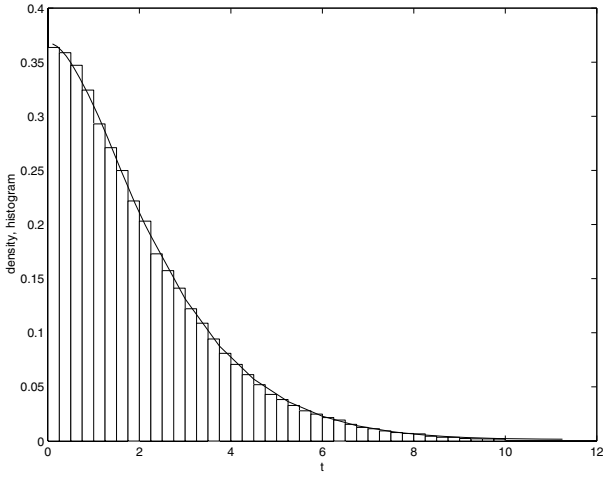


FIGURE 3. Density (numerical Laplace inversion) and Histogram (simulation) for the independent case.

4 CONCLUDING REMARKS

The paper provides an overview of historical developments about stochastic modeling degradation as compound point processes. The paper extends the classical results to more general case and illustrates that modern methods of computing the inverse of the Laplace transform can be applied to derive the distribution of cumulative damage in a more general setting.

Acknowledgments

The authors like to thank P. den Iseger for his help with and useful discussions about the numerical inversion of the Laplace Transforms. This work was done while the first author was visiting the University of Waterloo. We would also like to thank Jasper Anderluh for making the figures in this paper.

Bibliography

[1] A. Mercer and C. S. Smith. A random walk in which the steps occur randomly in time. *Biometrika*, 46(1-2):30–35, 1959.

[2] R. C. Morey. Some stochastic properties of a compound-renewal damage process. *Operations Research*, 14(5):902–908, 1966.

[3] W. L. Smith. Renewal theory and its ramifications. *Journal of the Royal Statistical Society. Series B*, 20(2):243–302, 1958.

[4] J. G. Shanthikumar and U. Sumita. General shock models associated with correlated renewal sequences. *Journal of Applied Probability*, 20:600–614, 1983.

[5] Sumita U. and J. G. Shanthikumar. A class of correlated cumulative shock models. *Advances in Applied Probability*, 17:347–366, 1983.

[6] R. E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing, Probability Models*. Holt, Rinehart and Winston, Inc, 1975.

[7] M. S. A. Hameed and F. Proschan. Nonstationary shock models. *Stochastic Processes and their Applications*, pages 383–404, 1973.

[8] A. W. Marshall and M. Shaked. Multivariate shock models for distributions with increasing hazard rate average. *Annals of Probability*, 7:343–358, 1979.

[9] H. C. Tijms. *A First Course in Stochastic Models*. John Wiley & Sons, New York, 2003.

[10] P. den Iseger. Numerical transform inversion using gaussian quadrature. *Probability in the Engineering and Informational Sciences*, 20:1–44, 2006.

[11] A. W. Marshall and I. Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, 62:30–44, 1967.