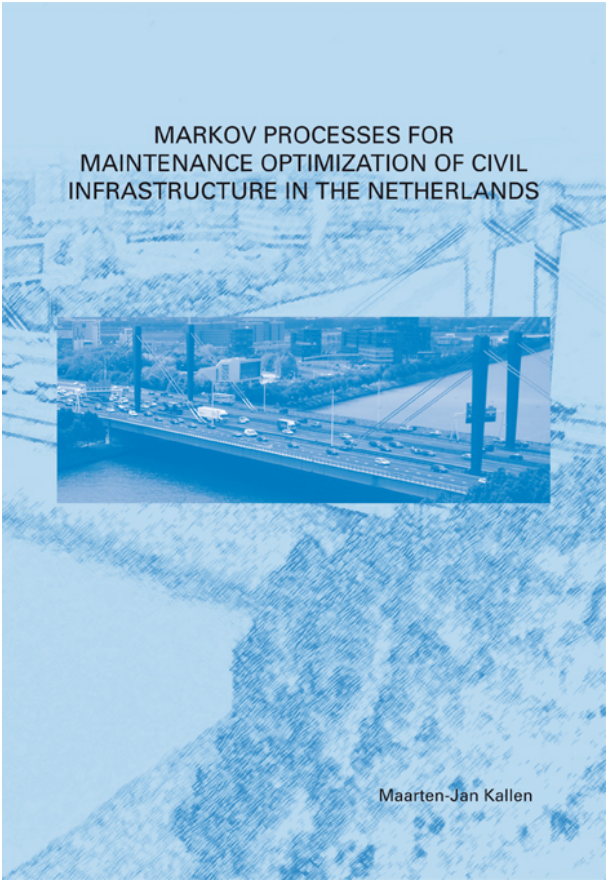


Cite this document as:

M.J. Kallen. *Markov processes for maintenance optimization of civil infrastructure in the Netherlands*. Ph.D. thesis, Delft University of Technology, Delft, 2007.

BibTEX entry:

```
@phdthesis{kallen2007phd,
    author  = {Kallen, M. J.},
    title   = {Markov processes for maintenance optimization of
civil infrastructure in the Netherlands},
    year    = {2007},
    school  = {Delft University of Technology},
    address = {Delft, Netherlands},
    ISBN    = {978-90-770051-29-0}
}
```

MARKOV PROCESSES FOR MAINTENANCE OPTIMIZATION OF CIVIL INFRASTRUCTURE IN THE NETHERLANDS

M.J. Kallen

MARKOV PROCESSES FOR
MAINTENANCE OPTIMIZATION OF CIVIL
INFRASTRUCTURE IN THE NETHERLANDS

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. Jacob Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 4 december 2007 om 10.00 uur
door Maarten-Jan KALLEN
wiskundig ingenieur
geboren te Creve Coeur, Verenigde Staten van Amerika.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. J.M. van Noortwijk

Samenstelling promotiecommissie:

Rector Magnificus,	Voorzitter
Prof. dr. ir. J.M. van Noortwijk,	Technische Universiteit Delft, promotor
Prof. dr. D.M. Frangopol,	Lehigh University
Prof. dr. A. Grall,	Université de Technologie de Troyes
Prof. dr. ir. R. Dekker,	Erasmus Universiteit Rotterdam
Prof. dr. T.A. Mazzuchi,	George Washington University
Prof. dr. ir. G. Jongbloed,	Technische Universiteit Delft
Dr. ir. A. van Beek,	Vereniging van Ondernemingen van Betonmortelfabrikanten in Nederland
Prof. dr. R.M. Cooke,	Technische Universiteit Delft, reservelid

Dit proefschrift is tot stand gekomen met ondersteuning van de Bouwdienst Rijkswaterstaat, HKV Lijn in water en de faculteit Electrotechniek, Wiskunde en Informatica van de Technische Universiteit Delft.

ISBN 978–90–770051–29–0

Cover design by Jan van Dijk, Dratex
On the cover: the 'Van Galecopper' bridge in Utrecht

Typeset with ConTEXt
Printed in The Netherlands

Contents

Summary

The Netherlands, like many countries in this world, face a challenging task in managing civil infrastructures. The management of vital infrastructures, like road bridges, is necessary to ensure their safe and reliable functioning. Various material restrictions, of which limited budgets are the most obvious example, require that the costs of inspections and maintenance must be balanced against their benefits.

A principal element of bridge management systems is the estimation of the uncertain rate of deterioration. This is usually done by using a suitable model and by using information gathered on-site. The primary source of information are visual inspections performed periodically. It is mainly due to the large number of bridges that these are not continuously monitored, but there are many other reasons why monitoring of all bridges is not practically feasible. The periodic nature of inspections creates specific requirements for the deterioration model.

This thesis proposes a statistical and probabilistic framework, which enables the decision maker to estimate the rate of deterioration and to quantify his uncertainty about this estimate. The framework consists of a continuous-time Markov process with a finite number of states to model the uncertain rate at which the quality of structures reduces over time. The parameters of the process are estimated using the method of maximum likelihood and the likelihood function is defined such that the dependence between the condition at two successive inspections is properly accounted for.

The results of the model show that it is applicable even if the data are subject to inspector interpretation error. Based on a data set of general conditions of bridges in the Netherlands, they are expected to require major renovation after approximately 45 to 50 years of service. This is roughly halfway the intended lifetime at design. The results also show significant uncertainty in the estimates, which is due to the large variability in a number of factors. These factors include the design of the structures, the quality of the construction material, the workmanship of the contractor, the influence of the weather, and the increasing intensity and weight of traffic.

A condition-based inspection model, specifically tailored to finite-state Markov processes, is proposed. It allows the decision maker to determine the time between inspections with the lowest expected average costs per year. The model, also known as the functional or marginal check-model, is based on renewal theory and therefore constitutes a life-cycle approach to the optimization of inspections and maintenance. In addition to this, a

complete chapter is devoted to determining the most computationally effective way of performing the necessary calculations in the deterioration and decision models. This ensures that analyses can be done almost instantly, even for very large numbers of structures.

The unified framework to deterioration modeling and decision making presented herein, contributes a quantitative approach to bridge management in the Netherlands and to infrastructure management in general. It can be applied to other fields of similar character like, for example, pavement and sewer system management.

Samenvatting

Nederland, zoals vele landen in deze wereld, staat voor een uitdagende taak in het beheer van civiele infrastructuur. Het beheer van belangrijke kunstwerken, zoals bruggen in het wegennet, is noodzakelijk om deze veilig en betrouwbaar te laten functioneren. Vanwege verschillende materiële restricties moeten de kosten van inspecties en onderhoud afgewogen worden tegen de baten. De meest voordehandliggende restrictie is die van een beperkt budget.

De schatting van de onzekere snelheid van veroudering is het belangrijkste element in een beheersysteem voor bruggen. Dit wordt gewoonlijk gedaan door gebruik te maken van een geschikt model en van gegevens die op lokatie zijn verzameld. De voornamelijkste bron van informatie zijn visuele inspecties die de beheerder periodiek laat uitvoeren. Het is vooral vanwege het grote aantal bruggen dat deze niet continu gemeten worden, maar er zijn veel meer redenen waarom dit in de praktijk niet haalbaar is. Het feit dat kunstwerken slechts periodiek geïnspecteerd worden, stelt bijzondere eisen aan het verouderingsmodel.

Dit proefschrift beschrijft een statistische en probabilistische aanpak die het de beheerder mogelijk maakt om de snelheid van veroudering te schatten en ook om zijn onzekerheid over deze schatting te kwantificeren. Het model bestaat uit een continue-tijd Markov proces met een eindig aantal toestanden om de onzekere snelheid van veroudering van kunstwerken over tijd te beschrijven. De parameters van dit model worden geschat door gebruik te maken van de methode van de grootste aannemelijkheid. De functie voor de aannemelijkheid is zodanig gedefinieerd dat deze de afhankelijkheid tussen twee opeenvolgende inspecties correct meeneemt.

De resultaten van het model tonen aan dat deze goed toepasbaar is, zelfs als de gegevens onderhavig zijn aan fouten die zijn gemaakt door de inspecteurs. Gebaseerd op een bestand van de algemene conditie van bruggen, hebben deze naar verwachting op een leeftijd van ongeveer 45 tot 50 jaar een grondige renovatie nodig. Dit is ruwweg halverwege de beoogde levensduur bij het ontwerp van een brug. De resultaten tonen ook een grote onzekerheid in de voorspelling, hetgeen komt door de grote variabiliteit in een aantal factoren. Voorbeelden van zulke factoren zijn het ontwerp van de kunstwerken, het vakmanschap van de aannemer, de kwaliteit van het materiaal, de invloed van het weer, en de toename in intensiteit en gewicht van het verkeer.

Een toestandsafhankelijk inspectiemodel, die geschikt is voor Markov processen met een eindig aantal toestanden, wordt gepresenteerd aan het einde van dit proefschrift. Het staat de beheerder toe om de tijd tussen inspecties te bepalen met de laagst verwachte gemiddelde kosten per jaar.

Dit model is gebaseerd op vernieuwingstheorie en beschouwd daarom de hele levenscyclus van het kunstwerk bij de optimalisatie van inspecties en onderhoud. Daarbovenop wordt een volledig hoofdstuk gewijd aan het bepalen van de meest efficiënte manier om de noodzakelijke berekeningen in het verouderings- en beslismodel uit te voeren. Dit zorgt ervoor dat de analyses in heel korte tijd uitgevoerd kunnen worden, zelfs voor een heel groot aantal kunstwerken.

Het complete concept voor het modelleren van veroudering en het nemen van beslissingen voor optimaal onderhoud, zoals deze in dit proefschrift beschreven worden, voegt een gedegen kwantitatieve aanpak toe aan het brugbeheer in Nederland en aan het beheer en onderhoud van civiele infrastructuur in het algemeen. Het kan toegepast worden in andere gebieden van een vergelijkbaar karakter, zoals bijvoorbeeld bij het beheer en onderhoud van asfaltering en riolering.

1

Introduction

In the year 2007, several bridges have made it into the news. Unfortunately, the news was not good. On August 1st, the I-35W Mississippi River Bridge in Minneapolis, Minnesota in the United States of America, collapsed during heavy traffic, killing 13 people. The images from the wreckage of the steel bridge were broadcast worldwide by television and internet. They showed the devastation resulting from the collapse of such a large structure.

On August 13th, an almost completed concrete bridge over the Tuo river near Fenghuang in the people's republic of China, collapsed killing 22 construction workers. Incidentally, the collapse occurred on the same day the Chinese government announced a plan to renovate over 6000 bridges which are known to be structurally unsafe.

In April, people in the Netherlands were confronted with the extremely rare announcement that a bridge would be closed for heavy traffic due to concerns about its load carrying capacity. This bridge, the 'Hollandse brug', is part of a highway connecting the cities of Amsterdam and Almere.

Bridges and viaducts play a vital role in today's transportation infrastructure and therefore are essential to today's economy. They are constructed and maintained in order to reliably fulfill this role, while also ensuring the safety of the passing traffic. However, most countries nowadays face an aging bridge stock and a strong increase in traffic. This makes bridge management a challenging problem, especially when budgets for maintenance are generally shrinking.

Aside from the loss of human life and the emotional impact of catastrophic incidents with bridges, the monetary costs can be extremely high as well. According to an estimate by the transportation industry in the Netherlands, the cost of the closure of the 'Hollandse brug' could run up to around €160 000 per day. The reconstruction of the Mississippi River Bridge was recently awarded for an amount of $238 million. The total costs of the bridge collapse, including the reconstruction, rescue efforts, and clean up, are estimated to be approximately $393 million by the Minnesota Department of Transportation.

There are many factors which make bridge management a complex problem. These include the occurrence of changes in construction methods and building codes over the years, the varying weight and intensity of traffic, the

large number of structures over a large area, the influence of the weather on the structure, and many more. These factors have one thing in common: they create uncertainty. The problem of bridge management is therefore a problem of decision-making under uncertainty. The uncertainty primarily lies in the lifetime of the structures. Over the years, many efforts have been made to better predict deterioration in bridges of all sorts in order to more effectively perform the maintenance of bridges.

The research presented in this thesis is aimed at modeling the rate of deterioration of bridges in the Netherlands. This is done by using information on the condition of bridges obtained by inspections performed between 1985 and 2004. A very large number of bridges in the Netherlands were constructed during the 1960's and 1970's. The design life of bridges is generally around 80 to 100 years. In the Netherlands, by experience, bridges require a major renovation approximately halfway their operational life. This means that the country is soon facing a wave of structures which are in need of renovation.

The remainder of this chapter provides a general introduction to bridge management and how maintenance modeling is used as part of this. There are many different types of mathematical models available, which can be used for the purpose of determining optimal maintenance policies. In Section 1.3, an overview is given of the current bridge management practices in the Netherlands and why one particular modeling approach, namely one that uses a finite-state Markov process for modeling the uncertain deterioration, is particularly suitable for application in the Netherlands.

1.1 BRIDGE MANAGEMENT

Bridge management is the general term used for the optimal planning of inspections and maintenance of road bridges. Most management systems will consider the bridges as a node in a road network in order to reduce unnecessary traffic obstructions and the number of maintenance actions. The necessity for bridge management systems (BMS) has grown in recent years. The construction of new bridges is slowing down and older bridges are starting to reach a critical age of about 40 years at which major maintenance and renovation work is necessary. Due to budget constraints, bridge owners are focusing increasingly on maintenance and repair instead of replacement.

Maintenance models are developed and used to balance the costs against the benefits (e.g., increased safety) of current and future maintenance and repair actions. A bridge maintenance system increases the scope of the analysis to the planning of maintenance for a network of bridges. Quoting Scherer and Glagola (1994): 'A BMS is defined as a rational and systematic

approach to organizing and carrying out all the activities related to managing a network of bridges'. The goal of this approach is the following: 'The objective of a BMS is to preserve the asset value of the infrastructure by optimizing costs over the lifespan of the bridges, while ensuring the safety of users and by offering a sufficient quality of service', which is quoted from Woodward et al. (2001).

Individual bridges are complex structures made up of multiple components and are constructed using several material types. Their structural behavior, the quality of the construction materials, and the intensity of traffic loads, are highly uncertain. Many models have been proposed to better predict the overall deterioration of bridges and to schedule inspections and maintenance such that costs and safety are optimally balanced.

1.2 MAINTENANCE MODELING

Maintenance, or the act of maintaining something, is defined as 'ensuring that physical assets continue to do what their users want them to do' by Moubray (1997). More formally, maintenance consists of any activity to restore or retain a functional unit in a specified state such that it is able to perform its required functions. The general goal of maintenance optimization may be formulated as 'the optimal execution of maintenance activities subject to one or more constraints'. In this definition, there are three aspects: what is optimal, what maintenance activities are available, and which constraints must be respected? An obvious constraint is a finite budget, which means that structures can not simply be replaced at any time and that maintenance can not be performed continuously. Constraints on the availability of construction material and qualified personnel may also create restrictions. There are many examples of maintenance, which may be small (like cleaning drainage holes) or large (like resurfacing the bridge deck), but inspections also represent an important activity. Inspections help gather information for making decisions and their results may influence future maintenance and therefore also the future condition of structures. This information is subsequently used in the last aspect to be discussed here, namely the aspect of optimization. Maintenance and inspections may be performed such that the costs are minimized, the reliability or availability maximized, the safety maximized, or that a combination of these is optimal in some way.

The challenging aspect of maintenance optimization, is that the state of a structure can not be accurately predicted throughout its lifetime. The time to reach a deficient condition is uncertain and varies strongly between different structures. This uncertainty is a result of many factors, including the quality of the construction material, the quality of the workmanship, the traffic intensity and the stress which is put onto the structure by heavy

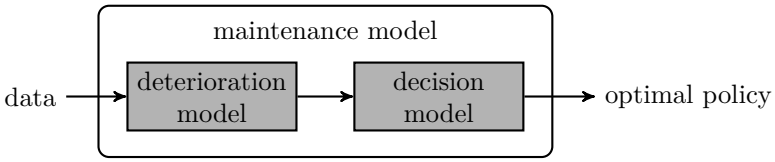


Figure 1.1: Simple representation of the two basic elements of a maintenance model: the deterioration and decision models.

loads. A natural variability exists due to for example differences in temperature, rainfall, wind and the presence of salt (e.g., in a maritime climate or in areas where frequent use of deicing salt is required). In order to make a sound decision on which maintenance policy is to be applied, the decision maker can use a model which represents an abstraction of reality and which quantifies the uncertainties involved in the degradation process. From this, it is obvious that such a model should be probabilistic in nature and not deterministic.

1.2.1 Elements of a maintenance optimization model

Maintenance models may be roughly divided in two parts: a deterioration model and a decision model. These two elements, as shown in Figure 1.1, are the basic parts in any maintenance model.

The deterioration model represents the abstraction of the actual degradation and the decision model uses the predicted deterioration to determine which maintenance policy is optimal. The decision model incorporates the decision criteria selected by the decision maker and uses information like, for example, costs of repair and the effectiveness of repair to calculate the optimal policy. Typical decision criteria are the inspection interval, condition thresholds for preventive repair, and the type of maintenance such as a complete renewal or a partial repair. Most of the variability and uncertainty is present in the deterioration model. The decision model may incorporate some uncertainty in the costs of repair, the effectiveness of lifetime extending maintenance, and in the discount rate, which is used to determine the value of investments and costs in the future.

The deterioration model may be supplied with data which is available to the decision maker. This data may include results from inspections in the form of condition and damage measurements, but it may also consist of estimates obtained using some form of expert judgment or a combination of these. As there are typically many structures in a network, the data is stored in a database to which new data is regularly added.

1.2.2 PHYSICAL VERSUS STATISTICAL APPROACH

Modeling the progress of deterioration over time can be done by using a physical or statistical approach. The physical approach entails the use of a model which attempts to exactly describe the deterioration process from a physical point of view. An example of such an approach is the use of Paris' law for modeling the growth of cracks in steel plates. Another example is the use of Fick's second law of diffusion for modeling the rate of penetration of chlorides in concrete. This model was fitted by Gaal (2004) to measurements of the chloride content in concrete samples taken from 81 bridges in the Netherlands.

A different approach to the problem of predicting deterioration based on historical data, is to assume that the data is generated by a mathematical model which does not try to emulate reality. Most commonly, this will be a probabilistic model which is fitted to the historical data by means of statistical estimation. An example of the statistical approach is the use of lifetime distributions fitted to lifetimes of bridges. This approach was used by van Noortwijk and Klatter (2004), where a Weibull distribution is fitted to ages of existing and demolished bridges in the Netherlands. The nature of this approach necessarily means that there is no 'true' model, but only models which fit better to the data compared to others; for example, see Lindsey (1996).

Other examples of the statistical approach are the application of stochastic processes like the gamma process and finite-state Markov processes. The gamma process has been used to model various types of degradation like, for example, thinning of steel walls of pressurized vessels and pipelines in Kallen and van Noortwijk (2005a) and the growth of scour holes in the sea-bed protection of a storm surge barrier in van Noortwijk et al. (1997). This process allows for a partial inclusion of physical knowledge by specifying the parameters in the expectation of the process, which is a power law function. Finite-state Markov processes, like Markov chains, have been used in the field of civil engineering to model uncertain deterioration in a number of areas like pavement, bridge, and sewer system management. One of the first examples is the Arizona pavement management system (Golabi et al., 1982), which inspired the Pontis bridge management system (Golabi and Shepard, 1997). More recently, Markov chains have been applied to sewer system and water pipeline deterioration. For examples, see Wirahadikusumah et al. (2001) and Micevski et al. (2002). A more complete overview of the application of both gamma processes and finite-state Markov processes is given by Frangopol et al. (2004). A specific review of the application of gamma processes in maintenance models is given by van Noortwijk (2007).

1.2.3 LIFE-CYCLE COSTING

During the lifetime of a structure, the condition is influenced by many external factors. The condition is also influenced by design decisions before construction, and maintenance actions after construction. Because every decision influences the timing and the nature of future decisions, it is important to take into account the effect of actions over the full lifetime of the structure. An important concept in infrastructure management is the concept of 'life-cycle costing'. All costs of construction, management and demolition must be taken into account by the decision maker. Due to the long design lives of bridges, the costs are usually discounted in time. Discounting is used to take into account the devaluation of money over time. The costs or rewards of future actions are therefore discounted towards their present value. Under the assumption that the costs of actions do not change over time, the result of discounting is that future actions are less costly. Money which is not spent now, can earn interest until it is needed for maintenance.

The timing of large scale repairs, like replacements, usually depends on the state of the structure. For example, in an age-based maintenance policy, a structure is repaired at fixed age intervals or when a necessity arises, whichever occurs first. If the structure has reached a predefined failure condition, it must be repaired or replaced. A common modeling approach is to use renewal theory which assumes that maintenance actions bring the structure to an as-good-as-new state. In this case, a repair is therefore equivalent to a replacement although it is usually not as expensive. The key idea behind renewal theory is that the timing of successive renewals is increasingly uncertain and that the probability of a renewal per unit of time will converge to a kind of average over the long run. As an example, the probability per year of a renewal using the Weibull lifetime distribution for concrete bridges in the Netherlands, as determined by van Noortwijk and Klatter (2004), is shown in Figure 1.2. Renewal theory supplies the decision maker with a number of convenient tools for the decision model in a maintenance model. A good theoretical presentation of renewal theory is given by Ross (1970).

1.3 BRIDGES AND THEIR INSPECTION IN THE NETHERLANDS

The Dutch Directorate General for Public Works and Water Management is responsible for the management of the national road infrastructure in the Netherlands. The Directorate General forms a part of the Ministry of Transport, Public Works and Water Management and consists of several specialist services. One of these is the Civil Engineering Division which is headquartered in Utrecht, the Netherlands. The Civil Engineering Division

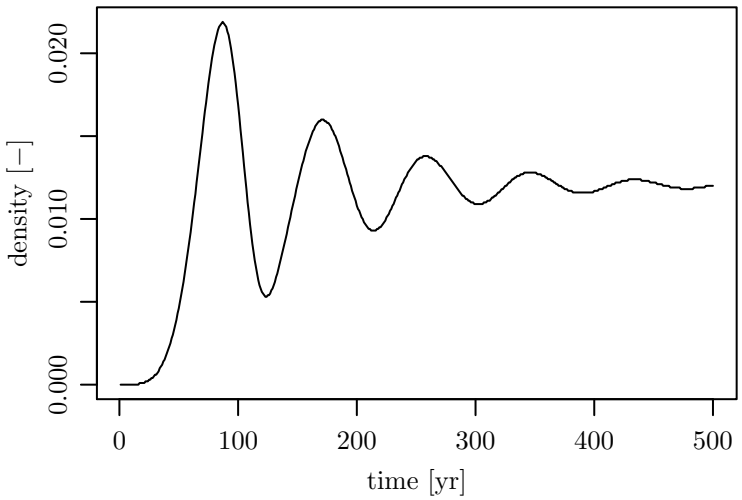


Figure 1.2: Renewal density using an estimated lifetime distribution for road bridges in the Netherlands.

'develops, builds, maintains, advises and co-ordinates infrastructural and hydraulic engineering structures that are of social importance'.

Since January 1st, 2006, the Directorate General has received the status of an agency within the ministry. The primary goal of this transformation is to apply a more businesslike approach to the execution of its tasks. As part of this new approach, the costs of business are weighed against the expected benefits. In general, the goal is to increase the accountability by clearly specifying what work is to be done, how it is to be done, and at what cost. Also, the satisfaction of the customer (i.e., the government and the people of the Netherlands) has become an even more important criterion. The commercial aspect also means that more engineering-like tasks (e.g., drawing and cost calculations) are outsourced to the market; that is, to commercial parties.

The national road network in the Netherlands consists of around 3200 kilometers of road, of which 2200 kilometers are highways. Within this network, there are approximately 3200 bridges, where the exact construction year is unknown for a little over 100 of these. Almost all bridges and viaducts are primarily concrete structures. About one hundred are mainly steel structures, aquaducts, or moveable bridges. The focus of this research is solely on concrete bridges, because form the largest group within the population. Also, the other structures can be considered as a fairly inhomogeneous group. Many of these structures are very unique in their design and construction.

Figure 1.3: Map of the Netherlands with the location of bridges which are managed by the Civil Engineering Division.

A map of the Netherlands with the location of the bridges is shown in Figure 1.3. A histogram of the construction years for concrete bridges in the Netherlands is presented in Figure 1.4. As can be observed in this figure, most bridges are currently between 30 and 40 years old. They have a life expectancy of about 80 to 100 years when designed. Due to increasing costs and a decrease in the availability of sufficient budgets for the construction and replacement of bridges, the focus is shifting more and more towards the efficient management of structures. The increased importance of infrastructure management has also resulted in the creation of a new 'maintenance and inspection' group within the Civil Engineering Division.

In the current inspection regime, large bridges (longer than 200 meters) are inspected every ten years, and smaller bridges every six years. Variable maintenance actions, which are defined as maintenance actions outside the long-term maintenance policy, are performed based on the condition of the

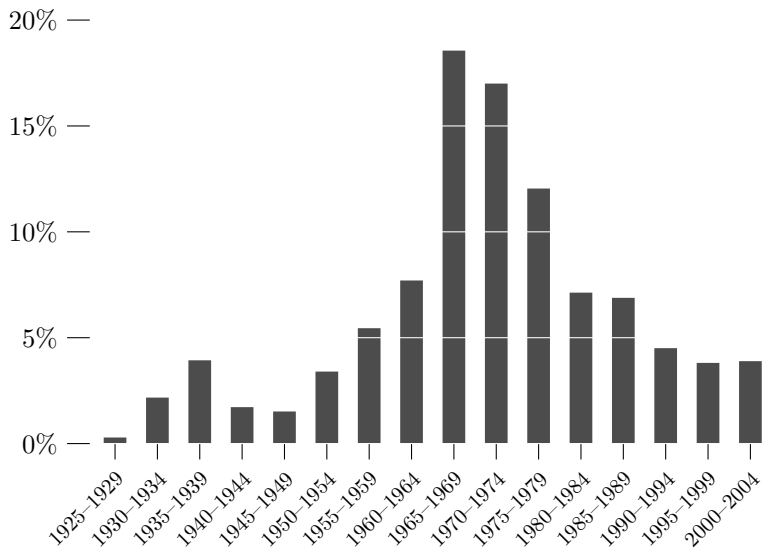


Figure 1.4: Histogram of construction years of concrete bridges in the Netherlands.

structures as observed during an inspection and routine maintenance is performed every year.

There are two types of inspections: functional and technical inspections. The functional inspections are performed more frequently and are primarily focused on analyzing the extent of individual damages or the state of materials. These functional inspections are usually performed by the regional office who is responsible for the structure. A technical inspection is a thorough analysis of the complete structure, aimed at registering the presence and severity of damages and at assessing the overall condition of the structure. The information gained from the technical inspections is used by the Civil Engineering Division for the purpose of managing the structures in the national road network. For this reason, and due to the fact that these inspections require specialized knowledge, the Civil Engineering Division is responsible for the planning and execution of these inspections.

The information gathered in a technical inspection is registered in an electronic database. The database includes the basic information of all structures in the Netherlands. This includes details like the location (province, community, highway, geographical coordinates, etc.), the size (length and width), if it is part of the highway or if it is located over the highway, the construction year, and which regional office of the Civil Engineering Division is responsible for regular inspections and maintenance.

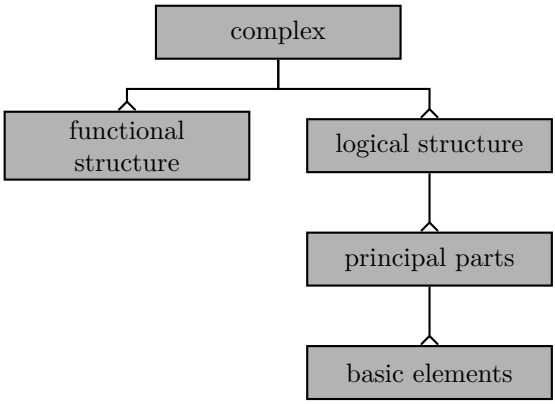


Figure 1.5: The relationship of structures and their elements in the Dutch bridge inspection database.

The largest objects in the database are the complexes, which may consist of one or more structures (e.g., different spans in a long bridge or two parallel bridges). Complexes are divided in two ways: functional or logical. The functional sectioning separates structures with different limits on traffic width, height and weight within the complex. This information is primarily used for the planning of special convoys which are particularly large or heavy. The logical separation is used for inspection purposes and separates the parts in the complex by the expertise which is required for the inspections. This means that, for example, all concrete, steel, moveable parts, and electrical components are considered as separate units for inspections. Each of these 'structures' is further divided in principal parts like for example, the superstructure of a bridge, and each principal part consists of one or more basic elements like, for example, the beams in the superstructure. A representation of this classification is shown in Figure 1.5.

The primary task of the inspector is to identify the damages and their location on the structures and to register these in the database. The damages are linked to the basic elements and their severity is quantified using the discrete condition scale shown in Table 1.1.

These condition states are the primary information on the extent of damages. More detailed information like, for example, the size of the damage, may be added to the database, but is generally not used in the planning and scheduling of maintenance. The system automatically assigns the highest (i.e., the worst) condition number of the basic elements to their parent primary component and the logical structure also receives the highest condition number of the primary components which it consists of. Because a minor component with serious damage will automatically lead to the structure as a whole to have a bad condition, the inspector is supposed to adjust

Code	State	Description
0	perfect	no damage
1	very good	damage initiation
2	good	minor damages
3	reasonable	multiple damages, possibly serious
4	mediocre	advanced damages, possibly grave
5	bad	damages threatening safety and or functionality
6	very bad	extreme danger

Table 1.1: Seven condition codes as used for the condition assessment of bridges in the Netherlands.

these assignments such that the overall condition number is representative for the structure.

1.4 AIM OF RESEARCH

The condition database as described in the previous inspection is used to gather information required for the planning and scheduling of maintenance and inspections. However, the historical development of condition numbers for bridges has not been used in a model for the estimation of the rate of deterioration. The classic approach for the deterioration model is to use finite-state Markov chains to model the uncertain rate of transitioning through the condition states. As indicated in Dekker (1996), Markov decision models are quite popular, mainly due to the fact that they are a natural candidate for condition data on a finite and discrete scale. This is also the reason why the gamma process is not considered in this research: it is more natural to apply the gamma process to modeling continuous deterioration. There are many publications which describe the use of Markov chains for deterioration modeling, some of which were mentioned in Section 1.2.2. Like 'Pontis' in the United States, a number of other countries have implemented, or at least experimented with, a bridge management system which is based on Markov chains. Examples in Europe are KUBAMS in Switzerland (Roelfstra et al., 2004), and PRISM in Portugal (Golabi and Pereira, 2003). In the Netherlands, there currently is no such system and the overall aim of this research is to develop a theoretical model and analyse its applicability using the Dutch bridge condition data.

A model may not be suitable for many reasons. For example, it may be too complicated to use, too inefficient to handle large amounts of data, it may be based on assumptions which are too restrictive, or it may not be able to deliver the necessary information for decision making. Even if there is a suitable model available, there may not be sufficient data or it may be

of too poor quality. Also, some models may be too expensive to implement, because they require the acquisition of very detailed information. The topic of this research is therefore also of a quite practical nature.

Finite-state Markov processes are a natural candidate for modeling the uncertain rate at which transitions through a discrete condition scale occur. Given this, the research is aimed at addressing the following issues:

a. can historical bridge condition data be extracted from the database in such a way that it can be used to estimate the parameters of the model?
b. what models have been proposed and applied before and what are their advantages and shortcomings?
c. what type of Markov process can be used and which procedure is most suitable for estimation of the model parameters?
d. how robust is the model and the estimation procedure to changes in the data?
e. how should the model be implemented, such that the calculations can be done efficiently and with sufficient accurracy?
f. how fast does the overall condition of concrete bridges deteriorate and how uncertain are the predictions given by the deterioration model?
g. does grouping of bridges based on selected characteristics result in significantly different parameters? In other words: is the bridge stock a heterogeneous population or are there noticable differences in the rate of deterioration?
h. is it possible and useful to include the variability or imperfection in the observations by inspectors into the model?
i. is there a suitable decision model for maintenance optimization and what information is required for the application of such a model?

1.5 READING GUIDE

The following chapter starts with a short overview of various aspects of finite-state Markov processes, which is suggested reading even for those familiar with this material as it introduces most of the notation used throughout this thesis. The rest of Chapter 2 contains an extensive review and evaluation of estimation procedures for Markov processes proposed in the past. It concludes with a short discussion on the applicability of the Markov property and on the use of semi-Markov processes.

Chapter 3 introduces a maximum likelihood estimation approach which constitutes a significant improvement over the past approaches. It is shown how perfect and imperfect inspections can be dealt with and how to test the significance of the influence of various characteristics of a structure on the outcome of the model. This chapter is mostly theoretical of nature.

The proposed maximum likelihood estimation is applied to the Dutch bridge condition data in Chapter 4. Various models are tested on data sets

of the overall bridge condition, superstructures and kerbs. This chapter presents the most important research results. One of the building blocks of this model is the transition probability function, which gives the probability of moving between any two condition states during a specified period of time. Chapter 7 describes the method of calculating this function, which is performed 'under the hood' and is therefore primarily of interest to those wishing to implement such a model.

The largest part of this thesis is concerned with the estimation of the deterioration process. Chapter 5 expands on this by considering a condition-based maintenance model which is particularly well suited to be used with finite-state Markov deterioration processes. Finally, conclusions and recommendations are given in Chapter 6.

In this thesis, the following notational conventions are used:

- matrices are denoted with boldface capital letters, like $\boldsymbol{P}$ and $\boldsymbol{Q}(t)$,
- $(\boldsymbol{P})_{ij}$ represents the (i, j) position or element of matrix $\boldsymbol{P}$,
- vectors are denoted with boldface letters, like $\boldsymbol{x}$ and $\boldsymbol{\theta}$,
- in matrix notation, all vectors are column vectors and their transpose is denoted with a prime, like $\boldsymbol{x}'$,
- indices are denoted with the letters i, j and k,
- random variables are denoted with capital letters, like T,
- the notations X_t, $X(t)$, Y_k and $Y(t_k)$ denote stochastic processes of various forms,
- the letters L and ℓ are reserved for the likelihood and log-likelihood respectively,
- the letters s, t and u represent time or age,
- the vector $\boldsymbol{\theta}$ represents a set of model parameters, and
- dimensions are given in square brackets, like [yr] for years and [-] for a value without a dimension.

2

Markov processes for bridge deterioration

Over the years, finite-state Markov processes have been applied quite frequently in the field of civil engineering. The main part of this chapter is formed by Section 2.3, which reviews several methods as used in applications towards civil infrastructure for the estimation of transition probabilities in Markov processes. For a better comprehension of this review, Section 2.1 first gives a short overview of the essential theory behind finite-state Markov processes and Section 2.2 describes the nature of bridge inspections and the type of data which follows from these inspections.

The chapter ends with some notes on typical issues, which have been raised over the past, relating to the application of Markov processes. These are: the validity of the Markov property and the use of semi-Markov processes to model aging. Here, aging is mathematically defined as an increasing probability of failure or transition to a lesser condition state as time progresses.

2.1 FINITE-STATE MARKOV PROCESSES

A finite-state Markov process is a stochastic process which describes the movement between a finite number of states and for which the Markov property holds. The Markov property says that, given the current state, the future state of the process is independent of the past states.

Let $\{X(t) \mid t \in \mathcal{T}\}$ represent the state of the process at time t and let X_k be the shorthand notation for $X(t_k)$, where $k = 0, 1, 2, \ldots$. According to the definition of a stochastic process, $X(t)$ is a random variable for every t in the $\mathcal{T}$. The set $\mathcal{T}$ is the index set of the process and because t represents time or age, the elements in this set are non-negative. Also, the process is assumed to always start at time $t_0 = 0$ and the set $\{t_k, k = 0, 1, 2, \ldots\}$ is an ordered set $t_0 < t_1 < t_2 < \cdots$. Using this notation, the Markov property formally states that

$$\begin{aligned}\Pr\{X_{k+1} = x_{k+1} \mid X_k = x_k, X_{k-1} = x_{k-1}, \ldots, X_1 = x_1, X_0 = x_0\} \\ = \Pr\{X_{k+1} = x_{k+1} \mid X_k = x_k\},\end{aligned}$$

where the set of possible states is taken to be finite and represented by a sequence of nonnegative integers: $x_k \in \mathcal{S} = \{0, 1, 2, \ldots, n\}$ for all k.

The probability of a transition taking place in Markov processes may depend on a number of time scales. As Commenges (1999) illustrates, there are three possible time scales: calendar time, age, and the time since the last transition. Calendar time is mostly of interest to epidemiologists. A simple example of a process depending on calendar time and age is the life of humans. It is known that in developed countries, mortality rates increase with age and decrease with calendar time. This means that as humans get older, they have a higher probability of dying and, on average, people get older now compared to those who lived in the middle ages. The age of a Markov process is defined as the time since the start of the process at t_0. If the transitions in a Markov process are independent of the age of the process, then the process is said to be stationary or time-homogeneous. The latter will be used from now on and a formal definition of time-homogeneity will be given in the following sections.

For civil infrastructures, the age of the process and the duration of stay in the current condition state are of most interest. A dependence on calendar time may for example be included to account for an increase (or decrease) in the quality of building materials or workmanship over the years.

The structure of Markov processes may be defined such that these are cumulative or progressive, which means that they proceed in one direction only. An example of a progressive Markov process is the pure-birth process; see for example Ross (2000). For modeling deterioration, finite-state Markov processes should posses at least two characteristics, namely:

1. the states represent conditions, therefore they must be strictly ordered, and
2. the process must progress monotonically through the condition states.

The process may also be sequential, such that the states are traversed one after the other and no state is skipped. A distinguishment is made between a discrete-time Markov process and a semi-Markov process. A discrete-time Markov process performs transitions on a discrete time grid, which is almost always equidistant. A semi-Markov process allows for transitions on a continuous time scale.

2.1.1 Discrete-time Markov processes

Let the index set be defined as $\mathcal{T} = \{0, 1, 2, \ldots\}$ and let $\{X_t, t \in \mathcal{T}\}$ be a Markov chain. For a time-homogeneous Markov chain, the probability of a transition between two states i and j per unit of time is defined by $p_{ij} = \Pr\{X_{t+1} = j \,|\, X_t = i\} = \Pr\{X_1 = j \,|\, X_0 = i\}$. The transition probabilities between all possible pairs (i, j), may be collected in the transition probability matrix

$$\boldsymbol{P} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0n} \\ p_{10} & p_{11} & \cdots & p_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n0} & p_{n1} & \cdots & p_{nn} \end{bmatrix}.$$

The matrix $\boldsymbol{P}$ is stochastic, which means that $0 \leq p_{ij} \leq 1$ for $i, j = 0, 1, 2, \ldots, n$ and $\sum_{j=0}^{n} p_{ij} = 1$ for all i. An alternative definition for the transition probabilities is given by $p_{ij} = \Pr\{P_i = j\}$, where P_i is the random variable describing the probability of the destination state if currently in state i. The transition probability matrix $\boldsymbol{P}$ not only defines the randomness of the process in time, but it also defines the structure of the model.

As an example of commonly used structures for the purpose of modeling deterioration, consider the transition probability matrices of a progressive and a sequential Markov chain:

$$\boldsymbol{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ 0 & p_{11} & p_{12} & p_{13} \\ 0 & 0 & p_{22} & p_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.1}$$

and

$$\boldsymbol{P} = \begin{bmatrix} 1 - p_{01} & p_{01} & 0 & 0 \\ 0 & 1 - p_{12} & p_{12} & 0 \\ 0 & 0 & 1 - p_{23} & p_{23} \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2.2}$$

Both examples have four successive condition states and their graphical representation is given in Figure 2.1. Note that, in these cases, state 4 is referred to as an 'absorbing' state and all other states are 'transient'.

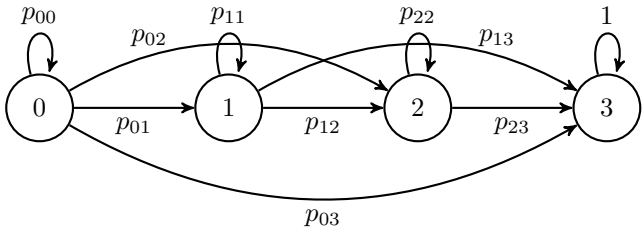
The Chapman-Kolmogorov equation, defined as

$$p_{ij}(m) = \sum_{k=0}^{n} p_{ik}(r) p_{kj}(m - r),$$

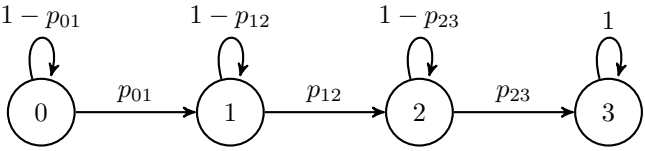
can be used to show that the probability $p_{ij}(m) = \Pr\{X_{t+m} = j \,|\, X_t = i\}$ of an m-step transition between any pair of states (i, j) may be calculated by multiplying the matrix $\boldsymbol{P}$ with itself m times and taking the (i, j)-th element, like $\boldsymbol{P}_{ij}(m) = \left(\boldsymbol{P}^m\right)_{ij}$.

2.1.2 Semi-Markov and continuous-time Markov processes

A semi-Markov process is an extension of a discrete-time Markov process in which a random time is added between transitions. Let J_0 be the state



(a)



(b)

FIGURE 2.1: Graphical representation of a progressive (a) and sequential (b) discrete-time Markov process.

of the process $\{X(t), t \geq 0\}$ at the beginning of the process and J_n, $n = 1, 2, \ldots$ the state of $X(t)$ after n transitions. The probability of the process moving into state j in an amount of time less than or equal to t, given that it just moved into state i, is defined as

$$Q_{ij}(t) = \Pr\{T_i \leq t, J_{n+1} = j \,|\, J_n = i\},$$

where T_i is the random waiting time in state i. This probability can be written as the product

$$Q_{ij}(t) = F_{ij}(t) p_{ij}, \tag{2.3}$$

where $p_{ij} = \Pr\{J_{n+1} = j \,|\, J_n = i\}$ is the transition probability function of the 'embedded' Markov chain $\{J_n, n = 0, 1, 2, \ldots\}$ and

$$F_{ij}(t) = \Pr\{T \leq t \,|\, J_{n+1} = j, J_n = i\}$$

represents the conditional probability of the random waiting time T given that the process moves into state j after previously having moved into state i. Equation (2.3) shows that transitions in a semi-Markov process have two stages: if the process just moved into state i, it first selects the next state j with probability p_{ij} and then waits a random time T according to $F_{ij}(t)$. The semi-Markov process $\{X(t), t \geq 0\}$ may be defined as $X(t) = J_{N(t)}$, where $N(t)$ is the total number of transitions during the interval $(0, t]$.

As for the discrete-time Markov process, it is interesting to know the probability of transitioning between a pair of states during a time interval of length $t \geq 0$. The transition probability function, defined as $p_{ij}(t) = \Pr\{X(t) = j \,|\, X(0) = i\}$ for time-homogeneous processes can be calculated by

$$p_{ij}(t) = \begin{cases} 1 - \sum_k \int_{x=0}^{t} [1 - p_{kj}(t-x)] \mathrm{d}Q_{jk}(x), & i = j \\ \sum_k \int_{x=0}^{t} p_{kj}(t-x) \mathrm{d}Q_{ik}(x), & i \neq j. \end{cases} \tag{2.4}$$

Obviously, $p_{ij}(0) = 0$ for $i \neq j$ and $p_{ii}(0) = 1$. This function is also referred to as the 'interval transition probability' by Howard (1971).

The name 'semi-Markov' stems from the fact that the process $X(t)$ is (in general) not Markovian for all t, because the distribution of the waiting time may not be a memoryless distribution. The Markovian property always holds at the times of the transitions. A special type of semi-Markov process arises when the waiting time is taken to be exponential; that is, when T_i has a cumulative distribution function given by

$$F_i(t) = 1 - \exp\{-\lambda_i t\} \tag{2.5}$$

with intensity $\lambda_i > 0$, and $p_{ii} = 0$ for all $i \in \mathcal{S}$. This implies that the process always moves to a different state and the waiting time is independent of which state it moves to. This type of semi-Markov process is referred to as a continuous-time Markov process, because it is Markovian for all $t \geq 0$. For continuous-time Markov processes, the transition probability function Equation (2.4) simplifies to

$$\boldsymbol{P}(t) = \exp\{\boldsymbol{Q}t\} = \sum_{k=0}^{\infty} \boldsymbol{Q}^k \frac{t^k}{k!}, \tag{2.6}$$

where $\boldsymbol{Q}$ is the transition intensity matrix with elements

$$q_{ij} = \begin{cases} -\lambda_i, & \text{if } i = j, \\ \lambda_i p_{ij}, & \text{if } i \neq j. \end{cases} \tag{2.7}$$

Note that $\sum_j^n q_{ij} = 0$ for all $i \in \mathcal{S}$. The function $\exp\{\boldsymbol{A}\}$, where $\boldsymbol{A}$ is a square matrix, is known as the 'matrix exponential'. An example equivalent to Figure 2.1 for continuous-time Markov processes is given in Figure 2.2.

2.1.3 First passage times and phase-type distributions

In maintenance and reliability modeling, one is often interested to know the time required to reach a particular state. For instance, if state j is defined

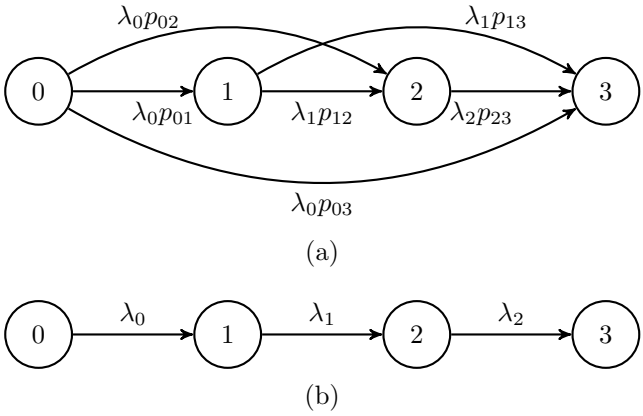


FIGURE 2.2: Graphical representation of a progressive (a) and sequential (b) continuous-time Markov process.

as a failed state and the process is currently in state i, the probability density of the first time of passage into state j for a discrete-time Markov process is defined as:

$$f_{ij}(t) = \Pr\{X_t = j, X_{t-1} \neq j, \ldots, X_1 \neq j \,|\, X_0 = i\},$$

with $t = 0, 1, 2, \ldots$ and $i \neq j$. This probability density function can be calculated using the recursive equation

$$f_{ij}(t) = \begin{cases} \sum_{k \neq j} p_{ik} f_{kj}(t-1), & t > 1, \\ p_{ij}, & t = 1. \end{cases}$$

For semi-Markov processes, the equivalent definition is

$$f_{ij}(n,t) = \Pr\{X(t) = j, X(s) \neq j \text{ for } \forall s \in (0,t) \text{ and } N(t) = n \,|\, X(0) = i\}$$

for $n = 0, 1, \ldots,$ and $t \geq 0$, which is a joint probability of the first passage time and the number of transitions required to first reach state j from i. This density can also be calculated recursively using the relation

$$f_{ij}(n,t) = \begin{cases} \sum_{k \neq j} \int_{s=0}^{t} f_{kj}(n-1, t-s)\, \mathrm{d}Q_{ik}(s), & n > 0 \text{ and } t > 0, \\ dQ_{ij}(t), & n = 1 \text{ and } t > 0, \\ 0, & n = 0 \text{ or } t = 0, \end{cases}$$

where $Q_{ij}(t)$ is given by Equation (2.3); see Howard (1971, p.733).

The time to reach an absorbing state in a finite-state Markov process has a so-called 'phase-type' probability distribution, because the process must pass through a finite number of phases before being halted by the absorbing state. Assume that the state set has $n+1$ states, that is $\mathcal{S} = \{0, 1, \ldots, n\}$,

and that the absorbing state is the last state in the process, then the transition intensity matrix $\boldsymbol{Q}$ may be divided as

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{r} \\ \boldsymbol{0}' & 0 \end{bmatrix},$$

where the $n \times n$ matrix $\boldsymbol{R}$ represents the transitions among the transient states, $\boldsymbol{r}$ is a column vector of length n with the intensities for transitions from the transient states into the absorbing state, and $\boldsymbol{0}'$ is the transpose of the column vector with all zeros. If the absorbing state is defined as being the failed state, the process is 'in service' or operative if it is in one of the transient states and so the probability of being in service in a time less than or equal to t is

$$\Pr\{X(t) < n\} = \boldsymbol{p}_0' \exp\{\boldsymbol{R}t\}\boldsymbol{1}.$$

Here, the row vector $\boldsymbol{p}_0'$ contains the probabilities of starting in one of the transient states and $\boldsymbol{1}$ is a column vector with all ones. The process is often assumed to start in state 0 with probability one, such that $\boldsymbol{p}_0 = \{1, 0, \ldots, 0\}$. The probability distribution of the time to failure is now simply given by

$$F(t) = 1 - \boldsymbol{p}_0' \exp\{\boldsymbol{R}t\}\boldsymbol{1} \tag{2.8}$$

with the probability density function $f(t) = \boldsymbol{p}_0' \exp\{\boldsymbol{R}t\}(-\boldsymbol{R} \cdot \boldsymbol{1})$. It should be clear that the above matrix analytic formulation can be used to determine the failure distribution for Markov processes with any arbitrary structure. Continuous-time Markov processes with a sequential structure like the example in Figure 2.2(b) have the following analytical solutions for the probability distribution of the time to failure:

- the Erlang distribution with probability density function

$$f(t) = \lambda \frac{(\lambda t)^{n-1}}{(n-1)!} \exp\{-\lambda t\}, \tag{2.9}$$

  if for all transient states $\lambda_i = \lambda$, and
- the hypoexponential distribution with probability density function

$$f(t) = \sum_{i=0}^{n-1} \left[\prod_{j \neq i} \frac{\lambda_j}{\lambda_j - \lambda_i} \right] \lambda_i \exp\{-\lambda_i t\}, \tag{2.10}$$

  with $\lambda_i \neq \lambda_j$ for $i \neq j$.

To summarize: if all waiting times have identical and independent exponential distributions, the time to absorption has an Erlang distribution (which is a special case of the gamma distribution) and if the waiting times

are exponential with strictly different intensity parameters, the time to absorption has a hypoexponential distribution. Both distributions may also be represented by Equation (2.8) with the appropriate intensity matrix $\boldsymbol{R}$.

First passage times have been used by Kallen and van Noortwijk (2005b) using a mixture of two Weibull probability distributions in a semi-Markov process fitted to bridge inspection data from the Netherlands. Phase-type distributions were formalized by Neuts (1981) using an algorithmic, or a matrix-analytic, approach. Their use is very common in queuing theory and they have also been used for modeling failure times, see e.g. Faddy (1995). Different names have been used for the distribution given in by Equation (2.10), like 'general gamma' or 'general Erlang' (Johnson et al., 1994), but 'hypoexponential' is most commonly used; for an example, see Ross (2000, p.253).

2.2 CHARACTERISTICS OF BRIDGE INSPECTION DATA

There are many ways to inspect a bridge. The quality and detail of information gathered during an inspection depends on the type of inspection which is applied. Inspections may be quantitative or qualitative. Quantitative inspections attempt to measure the physical properties of deterioration on structures. Examples are the measurement of chloride content in concrete and the sizing of cracks in steel. Qualitative inspection methods are generally subjective interpretations of the level of deterioration obtained by visual inspections. Most often, these type of inspection methods will result in the classification of the condition in a finite number of states.

Inspections are assumed to be periodic by definition and continuous measurements or observations of the condition of bridges are referred to as monitoring. Bridge 'health monitoring' is a rapidly growing field in the area of bridge management. Monitoring can, amongst others, be used to measure vibrations generated by traffic or measure contraction and expansion due to temperature changes.

This chapter deals solely with categorical inspection data from periodic observations, because quantitative inspections are not well suited for application on a large scale. Take for example the measurement of chloride content in concrete, which requires the drilling of core samples for analysis in a laboratory. The drilling of cilindrical test samples from bridges is time consuming and too costly to perform throughout the whole bridge network on a regular basis. Also, due to spatial variability, the results obtained from these samples are likely not to be representative for the whole structure.

The level of detail in data obtained from periodic observations can differ as well. Data may be kept for individual structures or may be pooled for a group of structures. Pooled data is usually referred to as 'aggregated data'. With this type of data, the number (or the proportion of the total number)

Code	State	Description
9	excellent	
8	very good	no problems noted.
7	good	some minor problems.
6	satisfactory	structural elements show some minor deterioration.
5	fair	all primary structural elements are sound; may have minor section loss, cracking, spalling or scour.
4	poor	advanced section loss, deterioration, spalling or scour.
3	serious	loss of section, deterioration, spalling or scour have seriously affected primary structural components. Local failures are possible. Fatigue cracks in steel or shear cracks in concrete may be present.
2	critical	advanced deterioration of primary structural elements. Fatigue cracks in steel or shear cracks in concrete may be present or scour may have removed substructure support.
1	imminent failure	major deterioration or section loss present in critical structural components or obvious vertical or horizontal movement affecting structure stability.
0	failed	beyond corrective action, out of service.

Table 2.1: Ten bridge condition codes as defined in FHWA (1995, p.38).

of structures in each condition state is known at successive points in time, but the transitions of individual structures are not known. If the condition history is known for each structure, the resulting data is known as 'panel data'. Finally, 'count data' is a special type of panel data where only the number of traversed states during an inspection interval is registered. In this case, the initial state and the target state are either not known, or not used by the decision maker.

Almost all discrete condition scales used in visual inspections represent the general or overall condition of a structure or one of its components. Therefore, different physical damage processes may lead to the same condition scale. This is something to keep in mind when modeling the condition of structures using a discrete and finite scale like those presented in Table 1.1 on page 11 and in Table 2.1. These rating schemes are typical for bridge management applications. A decreasing (or increasing) condition number is used to represent the decrease in the condition (or the increase in deterioration) of structures. The condition states and their identification are subject to personal interpretation and the scale is not necessarily equidistant, which means that the difference between 'excellent' and 'very good' is not necessarily the same as the distance between 'poor' and 'serious'. The following quote from FHWA (1995, page 37) illustrates quite well how these codes should be interpreted and used:

> "Condition codes are properly used when they provide an overall characterization of the general condition of the entire component being rated. Conversely, they are improperly used if they attempt to describe localized or nominally occurring instances of deterioration or disrepair. Correct assignment of a condition code must, therefore, consider both the severity of the deterioration or disrepair and the extent to which it is widespread throughout the component being rated."

The fact that discrete condition scales have no physical dimension has significant consequences for their application in maintenance optimization. Without information on the type of damage and the sizing of the damage, it is practically impossible to put a cost on repairs, replacements, or even on failures.

Inspections are generally assumed to be performed uniformly over time and over a group of structures, which means that some structures are not inspected more (or less) than others due to their state (or any other physical characteristic) or due to their age. This assumption is violated when certain structures, which are known to deteriorate faster than others, are inspected more often than others. In this case, the process of performing inspections depends on the rate of deterioration and is therefore not random. Another common situation in which this assumption may be violated is when structures are inspected immediately after a maintenance action in order to determine their 'new' condition. This introduces another issue which is of great influence in bridge inspections: maintenance. At the least, the rate of deterioration is slowed down by performing maintenance on structures and in most cases it will also result in an improved condition state.

The goal of performing regular periodic inspections is not only to ensure the safe operation of structures, but also to gain insight in the rate at which structures deteriorate. This insight may be used for optimizing the planning and scheduling of maintenance actions or the timing of subsequent inspections. As maintenance influences the rate of deterioration, it is imperative that this information is known to the modeler during estimation of the model parameters. Otherwise the results will not be representative of the real life situation. In fact, the estimated rate of deterioration will underestimate the actual rate when maintenance actions are ignored intentionally or unintentionally.

Another important issue involved with the estimation of deterioration rates of structures is censoring. Censoring arises when objects are not observed over their full lifetime. For structures, the process of deterioration is censored because bridge conditions are not continuously monitored and because inspection regimes are not the same over the full lifetime of the

structures. For example, the database used for registration of bridge inspection results in the Netherlands, has been in use since December 1985. Although structures were inspected before this time and the results were somehow registered, this information is not used in the decision making process because it is considered too old and it was obtained with a different inspection regime. So, in the case of the database in the Netherlands, the condition of structures built before 1985 is censored. The same holds for the end of the lifetime of structures. When the data set is used for analysis, most structures will not have reached the end of their service life. Besides this form of left- and right-censoring, there is also a kind of interval censoring in bridge inspection data. Periodic inspections of Markov deterioration processes reveal only current status data, which means that the decision maker knows only that one or more transitions have taken place between two inspections, but he does not know the times at which they occurred.

Finally, bridge condition data will never contain a set of observations uniformly distributed over all condition states. Even if inspections are assumed to be independent of state and age, civil infrastructures like bridges have long design lives and physical failures rarely occur. This means that in most data sets, there are many more observations of the better conditions relative to observations of poorer conditions.

2.3 REVIEW OF STATISTICAL MODELS AND ESTIMATION METHODS

The use of Markov processes with a finite number of states has become quite common in civil engineering applications. In order to fit the deterioration process to the available data, several statistical models and corresponding estimation methods have been proposed to determine the optimal values of the model parameters. The parameters in a Markov process are the transition probabilities or intensities, depending on whether a discrete- or continuous-time process is used. This review is divided in three parts with the division being based on the method of estimation: estimation methods other than maximum likelihood, maximum likelihood estimation, and less common methods like those using Bayesian statistics are also mentioned.

Classic statistical models are linear and generalized linear models or nonlinear models, which relate a response (or dependent) variable to one or more explanatory (or independent) variables using a linear or nonlinear function in the parameters. Generalized linear models form a broader class than the class of linear models. Besides linear models as a sub-class, generalized linear models include the binary probit (logit), ordered probit (logit), and Poisson models amongst others. This area of statistical analysis is commonly know as regression analysis.

2.3.1 METHODS OTHER THAN MAXIMUM LIKELIHOOD

This section deals with estimation methods which do not use the method of maximum likelihood. Traditionally, this form of regression uses the method of least squares or the method of 'least absolute deviation' to minimize the discrepancy between the model and the observations. A perfect fit is generally not possible due to the limitations of a simplifying model, nor is it desirable, as the model should be an abstraction of reality with the purpose of making some inference about the behaviour of the phenomenon being analyzed.

Two approaches are distinguished in this section: 1) minimizing the distance between the expectation of the condition state and the observations, and 2) minimizing the distance between the probability distribution of the condition states and their observed frequencies. In the first approach, the observations are the states of structures of various ages. The second approach uses the count (or the proportion) of structures in each state at various ages.

Regression using the state expectation

Fitting a Markov chain deterioration model by minimizing the distance between the observed states and the expectation of the model, is by far the most common approach found in the literature on infrastructure management. Assume that the condition of structures is modeled by the Markov chain $\{X(t), t = 0, 1, 2, \ldots\}$ and let $x_k(t)$ denote the k-th observation of a state at age t. In other words, the population of bridges is assumed to be homogeneous and for each t in a finite set of ages, there are one or more observations of the condition state. As the name suggests, the method of least squares minimizes the sum of squared differences between the observed state at age t and the expected state at the same age. This is formulated as follows:

$$\min_{p_{ij}} \sum_t \sum_k \big\{x_k(t) - \mathbb{E}X(t)\big\}^2, \tag{2.11}$$

under the constraints $0 \leq p_{ij} \leq 1$ and $\sum_j p_{ij} = 1$. The expectation of the Markov chain at time t is given by

$$\mathbb{E}X(t) = \sum_j j p_j(t),$$

where $p_j(t) = \Pr\{X(t) = j\}$ is the state distribution at time t and is defined as

$$p_j(t) = \sum_i \Pr\{X(t) = j \,|\, X(t-1) = i\} \Pr\{X(t-1) = i\}. \tag{2.12}$$

The model in Equation (2.11) deceptively looks like a linear model. However, it is a nonlinear model as the expectation of $X(t)$ is nonlinear as a function of the parameters, which are the transition probabilities.

The earliest references of the application of the least squares method in infrastructure management can be found in the area of pavement management. An overview of the early development is given by Carnahan et al. (1987) and Morcous (2006) also refers to Butt et al. (1987) as an example of the application in pavement management. Carnahan et al. (1987) and Morcous (2006) also discuss the use of the least absolute deviation regression, which minimizes the sum of the absolute value of the differences. A more recent application to pavement management is given by Abaza et al. (2004) and the regression onto the state expectation is also applied to sewer system management by Wirahadikusumah et al. (2001).

In Cesare et al. (1994), least squares minimization is applied to a slightly different model compared to the one presented in Equation (2.11). This approach consists of minimizing the weighted sum of squared differences between the observed proportion of states and the state distribution given by the process $X(t)$, which is given by

$$\min_{p_{ij}} \sum_t n(t) \sum_k \big\{ y_k(t) - p_k(t) \big\}^2, \tag{2.13}$$

where $n(t)$ is the number of observed states at time t, y_k is the observed proportion of structures in state k, and $p_k(t) = \Pr\{X(t) = k\}$ is the probability of the process $X(t)$ being in state k at time t. The weights $n(t)$ are used to assign more weight to those proportions which have been determined with more observations.

Probably the most significant objection against using these approaches is the fact that so much detail in the data is disregarded. The expectation of the Markov chain aggregates the historical development of the individual structures. Also, if only the expected condition at time t is available, the decision maker can not deduce the state distribution either. So even if successive observations of a single structure are available, the observations are treated as being independent and this is in contradiction with the assumption of the underlying Markovian structure. Another very strong objection against the formulation of the model in Equation (2.11), is the fact that the expectation of $X(t)$ depends on the definition of the condition scale. From this perspective, the model formulated in Equation (2.13) is much more appropriate.

Regression using the state distribution

The Pontis bridge management system uses the observed proportions of states for individual components. There are five sequential states, such that each component in a structure is assigned a vector $\boldsymbol{y} = \{y_1, \ldots, y_5\}$

after an inspection. This vector reflects that a proportion y_1 of the object in state 1, a proportion y_2 in state 2, etc. Obviously, it must hold that $\sum_{k=1}^{5} y_k = 1$. Assume that the condition of the component is modeled by a Markov chain $\{X(t), t = 0, 1, 2, \ldots\}$ and that at least two successive observations of the proportions, denoted by $\boldsymbol{y}(t-1)$ and $\boldsymbol{y}(t)$, are available. The probability of the proportions at time t is given by Equation (2.12) and because the observed proportions will generally not satisfy this relationship, an error term can be used to allow for the difference:

$$y_k(t) = \sum_{i=1}^{5} y_i(t-1)p_{ik} + e(t), \tag{2.14}$$

for $k = 1, \ldots, 5$. In Lee et al. (1970, Chapter 3) it is shown how this relationship can be used to obtain the classic estimator $\hat{p} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ with appropriately defined matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. This relatively easy relationship for the estimator is obtained by least squares optimization. Unfortunately, this approach does not explicitly take into account the constraints for the transition probabilities. The row sum constraint holds, but $0 \leq p_{ij}$ may be violated. An adjustment to the model is therefore required. Alternatively, the method of maximum likelihood could be used by choosing an appropriate probability distribution for the error term $e(t)$. Intuitively, the model in Equation (2.14) is quite appealing as it neatly incorporates the progressive nature of the Markov process. It does so by directly relating an observed condition state to the condition state at the previous inspection, using the transition probability.

The description of the multiple linear regression approach in AASHTO (2005) does not describe how the problem with the non-negativity constraint is accounted for. Another important constraint for the application of this approach is that there should be more observations than there are states. The Pontis system assumes that the states are sequential such that it is only possible to transition one state at a time. As the last state, the fifth state, is absorbing, there are just four transition probabilities to be estimated. The quality of the description of the methodology in AASHTO (2005) is quite poor and the methodology itself is faulty. The way the Pontis system attempts to combine transition probability matrices estimated from pairs of observations with different time intervals separating them, is a good example of this. First, the observation pairs are grouped in ten bins, where the first bin contains all pairs with 6 to 18 months separating them, the second bin contains all pairs that are observed 19 to 30 months apart from eachother, etc. Second, the transition probability is calculated for each bin. The one year transition probability $p_{ij} \equiv p_{ij}(1)$ is calculated using the transitions in the 6 to 18 month bin, the two year transition probability $p_{ij}(2)$ is calculated using the 19 to 30 month bin, and so on

up to the tenth bin. Third, the estimated transition probabilities for each bin are converted to a one year transition probability by the faulty relationship $p_{ij} = \sqrt[n]{p_{ij}(n)}$ for $i = j$, $p_{ij} = 1 - \sqrt[n]{p_{ij}(n)}$ for $i = j + 1$, and $p_{ij} = 0$ otherwise. Fourth, all converted transition probabilities are combined into the final estimated transition probability matrix by taking a weighted average of the ten transition probability matrices. The third and fourth steps are incorrect. A counter example for the third step is easily given. A move from state 1 to state 2 during two time periods can be achieved in two ways. The probability of this transition is therefore determined by $p_{12}(2) = p_{11}p_{12} + p_{12}p_{22}$. It is obvious that the square root of this probability is not equal to p_{12}.

2.3.2 MAXIMUM LIKELIHOOD METHODS

In most situations, the method of estimating model parameters by maximizing the likelihood of the observations, is a possible approach. This is the case is if, for example, the error term in the model is assigned a probability distribution, or if the parameters are probabilities themselves. A more detailed introduction to the concept of maximum likelihood estimation will be given in Chapter 3.

Poisson regression for continuous-time Markov processes

If an object has performed one or more transitions during the time between two periodic inspections, only the number of transitions and not the times of these transitions are known. In order to use count data to estimate transition probabilities, it is often assumed that the transitions are generated according to a Poisson process. A Poisson process is a stochastic process which models the random occurrence of events during a period of time. If the time between the occurrence of each event is exponentially distributed with parameter $\lambda > 0$, then the probability of n events occurring during a period with length $t \geq 0$ has a Poisson distribution. The probability density function of the Poisson distribution is given by

$$\Pr\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \tag{2.15}$$

with mean λt such that the expected number of events per unit time is $\mathbb{E}[N(1)] = \lambda$. If there are $m = 1, 2, \ldots$ independent observations (t_1, n_1), (t_2, n_2), …, (t_k, n_m), the likelihood of these observations is given by

$$\Pr\{N(t_1) = n_1, \ldots, N(t_m) = n_m\} = \prod_{k=1}^{m} \frac{(\lambda t_k)^{n_k}}{n_k!} e^{-\lambda t_k}. \tag{2.16}$$

The maximum likelihood estimator for λ is

$$\hat{\lambda} = \frac{\sum_{k=1}^{m} n_k}{\sum_{k=1}^{m} t_k}. \tag{2.17}$$

The term 'Poisson regression' stems from the fact that the parameter λ is often assumed to depend on one or more covariates in a multiplicative model: $\lambda = \exp\{\boldsymbol{\beta}'\boldsymbol{x}\}$, where $\boldsymbol{x}$ is a vector of covariates and $\boldsymbol{\beta}$ the vector of coefficients to be estimated. Poisson regression is therefore a generalized linear regression method with the logarithm as the link function; that is, $\log(\lambda) = \boldsymbol{\beta}'\boldsymbol{x}$, which is also known as a log-linear regression model.

For the application to bridge inspection data, the use of Poisson regression is restrictive in the sense that it requires substantial simplifications of the real life situation. The Poisson process counts the number of events and does not account for different types of events. The simplifying assumption is therefore that each event is the same, namely a transition to the next state after an exponential waiting time. The model is therefore necessarily sequential (because it is not possible to distinguish between different target states) and the waiting time in each state is the same. Another often mentioned limitation of the Poisson process is the fact that the variance of $N(t)$ is equal to its mean (and therefore increases when the mean increases), whereas the data may be more dispersed such that the variance should be greater than the mean.

Also, the simple likelihood given by Equation (2.16) and the estimator in Equation (2.17), which follows from it, do not account for the fact that the number of transitions is finite in the sequential model. Let $S_n = T_1 + T_2 + \cdots + T_n$ represent the random time required to perform n transitions. Knowing that the equivalence relationship $S_n \leq t \iff N(t) \geq n$ holds, it is possible to write $\Pr\{N(t) = n\} = \Pr\{S_n \leq t, S_{n+1} > t\}$. In words: the probability of exactly n transitions during time interval $(0, t]$ is equal to the joint probability that the n-th transition occurs before time t and the next transition occurs after time t. It is now quite easy to show that this approach does not work in the case of a finite process, like the Markov process considered here. Let the set of states be given by $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$ and $i, j \in \mathcal{S}$, then a transition from any $i \in \mathcal{S}$ to $j = 5$ during a period t requires special attention. Because there is no such thing as a 'next' transition in this case, the probability of the number of events during a period of length t is actually

$$\Pr\{N(t) = n\} = \begin{cases} \Pr\{S_n \leq t, S_{n+1} > t\}, & \text{if } j \neq 5, \\ \Pr\{S_n \leq t\}, & \text{if } j = 5, \end{cases} \tag{2.18}$$

where we again let $n = j - i$. Therefore, the probability of observing n transitions, in which the last transition was into the absorbing state 5, is given by

$$\Pr\{S_n \leq t\} = \Pr\{N(t) \geq n\} = \sum_{k=n}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = 1 - \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

This is the cumulative distribution function of the Erlang distribution for which the density function was given in Equation (2.9).

Given the result in Equation (2.18), we can reformulate the likelihood of all m observations in Equation (2.16) as

$$\Pr\{N(t_1) = n_1, \ldots, N(t_m) = n_m\} = \Big\{ \prod_{\forall\{k:j_k \neq 5\}} \frac{(\lambda t_k)^{n_k}}{(n_k)!} e^{-\lambda t_k} \Big\} \times \Big\{ \prod_{\forall\{k:j_k=5\}} \sum_{l=n_k}^{\infty} \frac{(\lambda t_k)^l}{l!} e^{-\lambda t_k} \Big\}. \tag{2.19}$$

The estimate $\hat{\lambda}$ must now be obtained by numerical methods.

The fact that the correct likelihood of the periodic observations is given by Equation (2.19) and not by Equation (2.16) gives an indication of the appropriateness of the nonparametric maximum likelihood estimator (NPMLE) suggested by Wellner and Zhang (2000) for periodically observed counting processes which are inhomogeneous. In these processes, the transition intensity depends on the age t of the process $X(t)$, such that $\lambda \equiv \lambda(t)$. The integrated intensity

$$\Lambda(t) = \int_{s=0}^{t} \lambda(s) ds \tag{2.20}$$

is also the mean function of the counting process defined by $\lambda(t)$: $\Lambda(t) = \mathbb{E}[N(t)]$. If a process is observed at successive times $0 < t_1 < t_2 < \ldots < t_m$ and each observation gives us the number of transitions since the last observation, denoted by $n_1, n_2, \ldots, n_m$, then the probability of these observations is given by

$$\Pr\{N(t_1) = n_1, N(t_2) = n_2, \ldots, N(t_m) = n_m\} = \prod_{k=1}^{m} \frac{(\Lambda(t_k) - \Lambda(t_{k-1}))^{n_k - n_{k-1}}}{(n_k - n_{k-1})!} \exp\left\{-(\Lambda(t_k) - \Lambda(t_{k-1}))\right\},$$

where $t_0 = 0$ and $n_0 = 0$. Again, this likelihood does not account for the finite number of transitions in a finite-state Markov process, therefore this model is not ideal for the purpose of estimating transition intensities in bridge deterioration models using finite-state Markov processes. It is noted that the NPMLE is a more general model which has the Poisson regression model as a special case.

Madanat and Wan Ibrahim (1995) used the likelihood in Equation (2.16) while acknowledging the fact that $N(t)$ is actually finite for the model under consideration. They mention the possibility of truncating the Poisson distribution as a possible correction, but assert that observations of the last state are very rare such that they do not influence the resulting estimator significantly. To account for possible overdispersion, the authors suggest the use of the negative binomial distribution instead of the Poisson distribution for the count of transitions. Compared to the Poisson distribution, which it has as a special case, the negative binomial distribution includes an extra parameter which allows the variance to be adjusted independently of the mean. This is a common approach to account for overdispersion, see Cameron and Travedi (1998, Chapter 4) for an example. In a Bayesian framework, the negative binomial distribution is derived by assuming that the intensity λ is gamma distributed.

Multinomial model for Markov chains

Assume that all structures are continuously monitored. For Markov chains this implies that each transition for every structure is observed. Let all observations be pooled by age $t \geq 0$ and let the set $\boldsymbol{N}_i(t) = \{N_{i1}(t), N_{i2}(t), \ldots, N_{in}(t)\}$ represent the random count of transitions to state $j = 1, \ldots, n$ from state i for all structures at age t. Because the deterioration process is continuously monitored, these counts are observed and are multinomially distributed for each state i. The probability of the observations $\boldsymbol{n}_i(t)$ at age t is given by the multinomial distribution with density function

$$f(\boldsymbol{n}_i(t)) = \Pr\{N_{i1}(t) = n_{i1}(t), \ldots, N_{in}(t) = n_{in}(t)\} = \frac{n_i(t-1)!}{\prod_{j=1}^{n} n_{ij}(t)!} \prod_{j=1}^{n} p_{ij}^{n_{ij}(t)},$$

with $n_i(t-1) = \sum_{j=1}^{n} n_{ij}(t)$ the total number of transitions out of state i and p_{ij} the transition probability from state i to state j. The likelihood of all observations is now simply given by

$$L(\boldsymbol{p}; \boldsymbol{n}) = \prod_{t=1} \prod_{i=1}^{n} f(\boldsymbol{n}_i(t)).$$

The maximum likelihood estimator for the transition probabilities is

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{n} n_{ij}},$$

where n_{ij} is the total number of observed transitions between states i and j over all ages of the structures.

This result was derived by Anderson and Goodman (1957) and Billingsley (1961). See also Lee et al. (1970) who refer to this type of data as 'micro data'. In the context of estimating bridge deterioration, Morcous (2006) referred to this method as the 'percentage prediction method'. The model as it is defined here, can not be used for bridge management, because it assumes that the deterioration process is continuously monitored. Since this is not the case, the exact count of transitions between condition states are not available. Application to the Dutch bridge condition data would require a substantial adjustment of this model in order to account for the censoring involved.

Probit and logit models for Markov chains

The binary probit and ordered probit models are linear regression models in which a continuous latent (unobservable) variable is observed to be in two (binary) or more (ordered) discrete categories. These models are appealing for the application in maintenance modeling, as the condition states are often assumed to be related to some underlying deterioration process which can not be measured directly.

Let the unobservable amount of deterioration be given by the random variable Y, then the probit model regresses this variable onto a linear model with standard normal errors ϵ:

$$Y = \boldsymbol{\beta}'\boldsymbol{x} + \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0,1), \tag{2.21}$$

where $\boldsymbol{x}$ is the vector of the explanatory variables (also referred to independent or exogeneous variables) with error ϵ. The row vector $\boldsymbol{\beta}$ contains the coefficients to be estimated and the first parameter, which is β_0, is usually taken as the intercept by setting $x_0 = 1$. Now let Z be a discrete random variable which represents the actual observed states, then the outcome for the binary probit model follows from

$$Z = \begin{cases} 0 & \text{if } Y \leq \tau, \\ 1 & \text{if } Y > \tau, \end{cases}$$

and for the ordered probit model with $n+1$ states it follows from

$$Z = \begin{cases} 0 & \text{if } Y \leq \tau_1, \\ 1 & \text{if } \tau_1 < Y \leq \tau_2, \\ \vdots & \\ n-1 & \text{if } \tau_{n-1} < Y \leq \tau_n, \\ n & \text{if } \tau_n < Y. \end{cases}$$

Because $Y \in (-\infty, \infty)$ for all $k = 1, 2, \ldots$, the thresholds τ and $\tau_i, i = 1, 2, \ldots$ for the state conditions must be located between $-\infty$ and ∞. Note

that the thresholds do not have to be equidistant. From these relationships, the probability of each observation can be determined. For the binary probit model this is simply $\Pr\{Z = 1 \,|\, \boldsymbol{x}\} = \Pr\{Y > \tau \,|\, \boldsymbol{x}\}$, where $\Pr\{Y > \tau\} = \Pr\{\boldsymbol{\beta}'\boldsymbol{x} + \epsilon > \tau\} = \Pr\{\epsilon > \tau - \boldsymbol{\beta}'\boldsymbol{x}\} = 1 - \Phi(\tau - \boldsymbol{\beta}'\boldsymbol{x})$. Here $\Phi(x)$ is the cumulative standard normal distribution function. The notation $\Pr\{Z = 1 \,|\, \boldsymbol{x}\} = \Phi(\boldsymbol{\beta}'\boldsymbol{x} - \tau)$ is also often used, which is equivalent as the normal distribution is symmetric with $\Phi(-x) = 1 - \Phi(x)$. Obviously $\Pr\{Z = 0 \,|\, \boldsymbol{x}\} = 1 - \Pr\{Z = 1 \,|\, \boldsymbol{x}\}$. Similarly, for the ordered probit model, the probabilities of observing each condition state is given by

$$\begin{aligned}
\Pr\{Z = 0 \,|\, \boldsymbol{x}\} &= \Phi(\tau_1 - \boldsymbol{\beta}'\boldsymbol{x}) \\
\Pr\{Z = 1 \,|\, \boldsymbol{x}\} &= \Phi(\tau_2 - \boldsymbol{\beta}'\boldsymbol{x}) - \Phi(\tau_1 - \boldsymbol{\beta}'\boldsymbol{x}) \\
&\vdots \\
\Pr\{Z = n-1 \,|\, \boldsymbol{x}\} &= \Phi(\tau_n - \boldsymbol{\beta}'\boldsymbol{x}) - \Phi(\tau_{n-1} - \boldsymbol{\beta}'\boldsymbol{x}) \\
\Pr\{Z = n \,|\, \boldsymbol{x}\} &= 1 - \Phi(\tau_n - \boldsymbol{\beta}'\boldsymbol{x})
\end{aligned} \tag{2.22}$$

Under the assumption that the observations are independent, the likelihood function for the coefficients $\boldsymbol{\beta}$ and the thresholds $\boldsymbol{\tau}$, given the observations $\boldsymbol{z}$ and the explanatory variables $\boldsymbol{x}_k$ is simply

$$L(\boldsymbol{\beta}, \boldsymbol{\tau} \,|\, \boldsymbol{z}) = \prod_k \Pr\{Z = z_k \,|\, \boldsymbol{x}_k\},$$

which can be maximized to estimate the unknown coefficients and thresholds. Before doing so, the model must be 'identified' by setting either the intercept β_0 or one of the thresholds τ_i equal to zero or some other constant. Fixing either the intercept or one of the thresholds will influence the other, but not the probability of the outcome z_k; see Long (1997, pp.122–123).

The logit model takes the same approach as the probit model, but assumes that the errors have a standard logistic distribution. The cumulative distribution function of the logistic distribution is

$$F(x) = \frac{1}{1 + e^{-(x-\mu)/s}}, \tag{2.23}$$

for $x \in (-\infty, \infty)$ and with mean μ and variance $\sigma^2 = (\pi s)^2/3$. The standard logistic distribution is symmetric and has $\mu = 0$ and $s = 1$, such that the variance is $\pi^2/3$. This is slightly more than the variance of the standard normal distribution. Because it is symmetric, it is possible to derive the same probability of the outcomes as in Equation (2.22) with $F(x)$ from Equation (2.23) instead of the cumulative standard normal distribution $\Phi(x)$.

The ordered probit model is generally not used to estimate transition probabilities in a Markov process, but Madanat et al. (1995) made some assumptions in order to apply this method to a Markov chain for modeling bridge deterioration. The first assumption is that the Markov chain is progressive with a transition probability matrix like in Equation (2.1). Then, the observations z_k are assumed to be the number of transitions between two consecutive inspections: $z_k = j - i$. A different Z is defined for each row except for the last in the transition probability matrix, thus allowing for different deterioration mechanisms in each (transient) state. Therefore, $\Pr\{Z_i = z\} = \Pr\{X(1) = i + z \,|\, X(0) = i\}$ for $i = 0, 1, \ldots, n-1$ and the authors introduce additional notation to allow the transition probabilities to be estimated for each individual bridge. Also, Madanat et al. (1995) use a log-linear model instead of the linear model in Equation (2.21) to ensure that the unobserved condition is non-negative: $\log(y_k) = \boldsymbol{\beta}'\boldsymbol{x}_k + \epsilon_k$. Then, the latent variable Y has a lognormal distribution with support $[0, \infty)$ and the thresholds for the condition states are also within this range. The software that the authors have used for the estimation, identified the model by setting the first threshold equal to 0, which corresponds to setting $\log(\tau_1) = 0 \Rightarrow \tau_1 = 1$. This model was later extended by Madanat et al. (1997) to a random effects model by the inclusion of another error term to reflect the differences (heterogeneity) between structures.

The approach suggested by Madanat et al. (1995), which was later applied by Baik et al. (2006) to the problem of modeling deterioration of wastewater systems, has a number of shortcomings. In what they see as an advantage, the option to estimate transition probability matrices for individual bridges requires a significant amount of inspection data and the suggested averaging of transition probabilities to obtain transition matrices for groups of bridges is faulty. Transition probabilities for groups of bridges should be directly estimated using the inspection data from all bridges within the group and not by averaging the transition probabilities of the individual bridges. A more fundamental shortcoming is related to the dependence of transition probabilities on bridge ages. The authors state that "the transition probabilities are explicitly ... nonstationary", because they are a function of time or the age of the bridge. The truth is that the aspect of time is included as an explanatory variable in the linear model and it is used to estimate a transition probability matrix of a stationary Markov chain. For example, take

$$Y(t) = \beta_0 + \beta_1 t + R, \text{ with } R \sim \mathcal{N}(0, 1) \tag{2.24}$$

as a simple model to describe the uncertainty in deterioration over time t. The probability of no transition between time $t_0 = 0$ and the first inspection at time t_1 is given by the probability that the amount of deterioration at time t_1 has not exceeded the first threshold τ_1: $\Pr\{Z = 0 \,|\, t_1\} =$

$\Pr\{Y(t_1) \leq \tau_1\}$. However, this probability is taken as the probability p_{00} of no transition out of the initial state 0 during a unit time. Subsequently, p_{01} is the probability that the amount of deterioration is somewhere between τ_1 and τ_2, p_{02} that it is somewhere between τ_2 and τ_3, etc. Therefore, each transition probability in a row of the transition probability matrix is related to a different age, but they are used in a transition matrix for a single unit time which is used to model transitions at all ages.

Although the inclusion of an unobserved continuous deterioration mechanism may seem to be attractive, the example model in Equation (2.24) shows that the linear model with Gaussian errors is really too restrictive to model uncertain deterioration. Since the expected amount of deterioration at $t = 0$ is equal to the intercept β_0 in the example model, it makes sense to indentify the model by setting $\beta_0 = 0$ instead of fixing one of the state condition thresholds, as the object is expected to be in a perfect state at the beginning of its service life. The coefficient β_1 can be interpreted as the rate of deterioration per unit time. Because the uncertainty in the amount of deterioration is added to the model as a random error, the variance in the amount of deterioration is constant over time. Also, the error is assumed to be standard normal such that the standard deviation is always equal to one. The choice for a unit variance is a convention, because the variance may be chosen arbitrarily. The choice for $\mu = \mathbb{E}[R]$ and σ^2 is part of the model identification. It affects the coefficients $\boldsymbol{\beta}$ but not the probability of the outcome. Together with the fact that only linear deterioration can be modeled, these characteristics of the model in Equation (2.24) make it unattractive for modeling uncertain deterioration over time. It is therefore advised not to include time or age as an explanatory variable in the regression model.

Under the assumption that the Markov chain is sequential with a transition matrix similar to the one shown in Equation (2.2), Bulusu and Sinha (1997) proposed to use a binary probit model for fitting the Markov chain to inspection data. A restrictive requirement for the application of this model is that only one transition occurs during the time between two successive inspections. The problems previously described for the probit model suggested by Madanat et al. (1995) are further aggrevated by the inclusion of a binary random variable in the linear model from Equation (2.21), which equals one if a transition took place in the previous inspection interval and zero otherwise. This attempt at incorporating time dependence into the model, directly violates the Markov property, which must hold if a Markov chain is used.

Similar to Bulusu and Sinha (1997), Yang et al. (2005) apply the same model, but without the extra binary variable and the errors R are assumed to have a logistic distribution as in Equation (2.23). This is therefore a binary logit model as described earlier. The authors refer to this approach

as 'logistic regression', although this terminology is also used by some to refer to regression with a log-linear model.

Non-parametric models

On page 31, the nonparametric maximum likelihood estimator for counting processes was already shortly mentioned. Parametric models have a finite number of parameters, whereas nonparametric models do not. A typical example of a parametric model is the continuous-time Markov process with constant transition intensities; see Equation (2.7). Another example is a inhomogeneous Markov process with transition intensities which depend on the age of the structure: $q_{ij}(t)$. A parametric approach assumes a continuous (or smooth) function with a finite number of parameters for the transition intensities. The nonparametric estimate $\hat{\lambda}(t)$ for counting processes proposed by Wellner and Zhang (2000) is a piecewise continuous function of t which is constant between the occurrence times of transitions.

Non-parametric models are typically used in the analysis of life data, which is a field of research known as 'survival analysis'. They are most appropriate for analysis of past events and do not lend themselves very well for future predictions. If one would want to make decisions on future performance, based on past experience, the nonparametric model must be smoothed in order to determine a future trend. A minimum requirement of the nonparametric estimators is that at least some events are observed. In the case of transitions between conditions, this means that the actual times of transitions must be observed. Various forms of censoring may be incorporated into the estimators, but a minimum number of actual observations are absolutely necessary.

DeStefano and Grivas (1998) discuss the application of a nonparametric estimator for the waiting time distribution in a semi-Markov process. Several assumptions are made: condition states transition only to the next state and the waiting time distribution is uniform with parameters $t_{\min}$ and $t_{\max}$. The somewhat unusual choice for a uniform distribution is due to the assumption that no specific knowledge about the waiting time is available or can be obtained. The authors concede that periodic inspections will never supply the decision maker with exact transition times, therefore they assume that if a transition occurred between two inspections, it occurred halfway the inspection interval. As the bridges are assumed to be inspected twice a year, this transition time is assumed to be sufficiently accurate. Also, it is implicitly assumed that no more than a single transition can take place between two inspections. With this critical assumption it is then possible to use the well known Kaplan-Meier estimator for the survival function. Although a Weibull probability distribution for the waiting time would have been more appropriate (to model increasing transition rates), it is primarily the subjective selection of 'virtual' transition times which

make the model proposed by DeStefano and Grivas (1998) unattractive. Detailed inspections can not be performed throughout a large network at a sufficiently high frequency to ensure that no significant errors result from the proposed assumption.

Finally, a nonparametric estimator known as the 'Aalen-Johansen' estimator is mentioned. Aalen and Johansen (1978) proposed this estimator for inhomogeneous Markov chains under censoring. Like the Kaplan-Meier estimator, it is a so-called 'product-limit' estimator which uses the concept of the product integral; see Section 7.2.2 in the appendix. The estimator has not been applied in the context of deterioration of civil infrastructures. Like all other nonparametric estimation models, this estimator can not be used if only instantaneous condition measurements are available. Aalen and Johansen (1978) discuss a general censoring process which takes on the values zero and one. This process must be integrable, meaning that it should be equal to one for a measurable amount of time. In other words: the structure must be observed during a longer period of time at least once in its lifetime. As bridge inspections are assumed to be instantaneous, this approach is not viable for the application in bridge management.

2.3.3 BAYESIAN METHODS

Instead of the maximum likelihood approach in the section entitled "Multinomial model for Markov chains" on page 32, it is possible to use a Bayesian approach. The multinomial distribution of the observed count of transitions out of each state is now the likelihood of the observations given the model parameters. In the Bayesian framework, the model parameters, being the transition probabilities, are given a prior distribution which reflects the modeler's belief in the value each p_{ij} and his uncertainty about these values. Let $g_i(p_{i1}, \ldots, p_{in})$ represent the prior probability distribution for the transition probabilities out of state i, then the natural candidate for this distribution is the Dirichlet distribution. The Dirichlet distribution, or multivariate beta as it is referred to by Lee et al. (1970), defined as

$$g(p_{i1}, \ldots, p_{in}) = \frac{\Gamma\left(\sum_{j=1}^{n} \alpha_{ij}\right)}{\prod_{j=1}^{n} \Gamma(\alpha_{ij})} \prod_{j=1}^{n} p_{ij}^{\alpha_{ij}-1}, \tag{2.25}$$

with parameters $\alpha_{ij} > 0, i, j = 1, \ldots, n$ and the gamma function

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} \mathrm{d}t, \text{ for } a > 0,$$

is a conjugate prior for the multinomial distribution and it restricts the random variables to $0 \leq p_{ij} \leq 1$ and to $\sum_{j=1}^{n} p_{ij} = 1$. When using a conjugate prior distribution, the posterior distribution $g(\boldsymbol{p} \,|\, \boldsymbol{n}_{ij}(t)) \propto$

$f_i(\boldsymbol{n}_{ij}(t) \mid \boldsymbol{p})g(p_{i1}, \ldots, p_{in})$, where $\boldsymbol{p}_i = \{p_{i1}, \ldots, p_{in}\}$ and $\boldsymbol{n}_i(t) = \{n_{i1}(t), \ldots, n_{in}(t)\}$, belongs to the same family. The parameters of the posterior distribution are simply the count of transitions added to the parameters of the prior distribution: $\alpha_{ij} + n_{ij}$. The posterior estimate for the transitions probabilities is

$$\hat{p}_{ij} = \mathbb{E}(p_{ij} \mid n_{i1}, \ldots, n_{in}) = \frac{\alpha_{ij} + n_{ij}}{\sum_{j=1}^{n} \alpha_{ij} + n_{ij}}.$$

Bulusu and Sinha (1997) compare the application of the Bayesian approach as described here to the application of a binary probit model to bridge deterioration modeling. For a more elaborate derivation, see Lee et al. (1970). Neither references discuss how the parameters α_{ij} of the prior distribution can be elicited by experts. A sensible approach is proposed by van Noortwijk et al. (1992) where experts give their initial estimates for the transition probabilities p_{ij} as percentages. First, they estimate the number n^*_{ij} of transitions from state i to j per unit of time. This estimate results in an expected value for the transition probability p_{ij} by setting $p^*_{ij} = n^*_{ij}/n_i$, where the total number of transitions out of state i is set to $n_i = 100$. Next, the parameters for the prior distribution in Equation (2.25) follow by setting the expectation for P_{ij}, which is $\mathbb{E}[P_{ij}] = \alpha_{ij}/\alpha_0$ with $\alpha_0 = \sum_j^n \alpha_{ij}$ equal to the estimate p^*_{ij}, such that

$$p^*_{ij} = \frac{\alpha_{ij}}{\alpha_0} \Longrightarrow \alpha_{ij} = \alpha_0 p^*_{ij}.$$

The parameter α_0 controls the variance of the estimator and thus reflects the strength of belief in the inspector's estimate. It is determined separately. Aside from this, special care should also be taken with weighting the opinion of multiple experts.

For the estimation of transition probabilities in a Markov model for storm waterpipes, Micevski et al. (2002) apply a different Bayesian approach. First, a non-informative (uniform) prior distribution is used for the transition probabilities. Second, the likelihood of the observed states is given by

$$L(\boldsymbol{n}(t) \mid \boldsymbol{\theta}) = \prod_t \prod_j p_j(t)^{n_j(t)},$$

where $p_j(t) = \Pr\{X(t) = j\}$ is the state distribution as defined in Equation (2.12) and $n_j(t)$ is the number of states observed to be in state j at time or age t. The prior distribution is not conjugate, therefore the authors use a Markov chain Monte Carlo method to sample the posterior distribution over the unknown transition probabilities. Notice that the observation data is in the form of state counts, which is similar to the aggregate data

used in the regression based models discussed in the section entitled "Regression using the state distribution" on page 27. This approach therefore suffers from the same shortcoming that the observed states are treated as conditionally independent. There is no probabilistic link between two consecutive inspections of a structure, even though this is the basic assumption of the underlying Markov chain.

2.4 TESTING THE MARKOV PROPERTY

There exist statistical tests to verify if the Markov property holds for the observed data. In general, all tests require full observations of the Markov processes and a sufficiently large number of observations to ensure accurate results. Unfortunately, both requirements are often not met. This poses a problem for the decision maker, who has assumed that the data is Markovian and wants to verify if his assumption is valid. For Markov chains where each transition is observed, there are a number of tests for the Markov property and for time dependence. One of the earliest references for these type of tests is Anderson and Goodman (1957). A more recent discussion in the context of economic analysis can be found in Bickenbach and Bode (2003).

A simplified test, equivalent to the tests described by Anderson and Goodman (1957), for the Markov property in bridge inspection data is presented in Scherer and Glagola (1994). Basically, the authors test the statistical significance of the difference between the probability of a sequence of three states $\{i, j, k\}$ and the probability of the sequence $\{j, k\}$ by use of a contingency table. Let

$$p_{ijk} = \Pr\{X(t_3) = k \,|\, X(t_2) = j, X(t_1) = i\}, \tag{2.26}$$

then the Markov property assumes that $p_{0jk} = p_{1jk} = p_{2jk} = \cdots = p_{jk}$. The number of observations for each sequence can be placed in a contingency table in order to calculate the value of the Chi-square test. Although not all possible sequences could be tested (due to lack of data), those sequences that were tested by Scherer and Glagola (1994) did not result in a rejection of the null hypothesis, which is the hypothesis that the (first order) Markov property holds.

Whether or not a test for the Markov property can effectively be performed largely dependends on the type of data which is available to the decision maker. As explained in the introduction to this chapter, observed bridge condition states may be aggregated, which makes testing the Markov property practically impossible. In the Netherlands, bridges are inspected throughout the year and at different intervals, which results in panel data. Although the individual state histories are known to the decision maker,

the times between inspections may vary significantly. Therefore, it is possible to count the number of sequences $\{i, j, k\}$ in the data but the states in this sequence may be observed at very different times. A simple count of sequences therefore ignores the fact that the observation times t_1, t_2 and t_3 in Equation (2.26) are almost always different. This makes the validity of such tests questionable.

Unless the transitions of a Markov process are directly observed and a large number of observations are available, it must be concluded that testing the hypothesis of the Markov property is not practically feasible. This conclusion was also drawn by Kay (1986), who proposed a possible solution using interpolation to recreate transition times, but conceded that this is a rudimentary workaround.

2.5 USING SEMI-MARKOV PROCESSES

Markov chains and continuous-time Markov processes are most often applied due to their computational tractability. However, concerns are sometimes raised about the constant transition intensity in each state. Because deterioration is at least partially or completely due to aging, some argue that it would be appropriate for the model to include aging. The most common probability distribution for the waiting time in each condition state, which enables the inclusion of aging, is the Weibull distribution. The cumulative distribution function for the Weibull distribution is given by

$$F(x) = 1 - \exp\{-(\lambda x)^{\beta}\} \tag{2.27}$$

with scale parameter $\lambda > 0$ and shape parameter $\beta > 0$. The hazard rate, defined by

$$h(x) = \frac{f(x)}{1 - F(x)},$$

for the Weibull distribution is $\beta\lambda(\lambda x)^{\beta-1}$. If $\beta = 1$, then the hazard function is constant and the Weibull distribution reduces to the exponential distribution defined in Equation (2.5). If $\beta > 1$ the hazard rate is increasing and therefore the probability of a transition out of the current state in the immediate future increases as the length of stay increases.

Although this approach may seem more appropriate from a physical point of view, the absense of the memoryless property results in a model which is difficult to work with. This is especially true when the deterioration process is not continuously monitored and only panel data is available. An example is the calculation of the time to failure in Kleiner (2001), which is given by a sum of Weibull distributions. As there is no analytical solution for the sum of Weibull distributed random variables, the author uses Monte Carlo simulation to obtain a numerical approximation. In the same paper, the

author concedes that there is insufficient data to estimate the parameters of the waiting times and proposes that these be determined by expert judgment. Mishalani and Madanat (2002) suggest to use a maximum likelihood approach which takes into account left- and right-censoring in the waiting time. However, their approach assumes that there are only two states and that the time of (re)construction is known. Each inspection therefore results in one of two possible observations: the transition has taken place before the inspection or it has not yet taken place. Another feature of the approach in Mishalani and Madanat (2002) is that the Weibull distributed waiting time is used to determine transition probabilities in a Markov chain which is necessarily nonstationary.

Another approach based on maximizing the likelihood of the observations is presented by Black et al. (2005b) and compared to Markov chains and the delay time model in Black et al. (2005a) using a case study. This approach is quite similar to the regression methods which use the state distribution as previously discussed in the section "Regression using the state distribution" on page 27, only now the waiting times are represented by Weibull random variables. This approach therefore suffers from the same shortcomings, like the fact that successive observations are essentially treated as being independent. An important assumption is that the construction year and the initial state distribution are known, such that the probability of being in an observed state can be used to calculate the likelihood of the observations. The reason why there is a common perception that the probability of a transition should increase as the time spent in a state increases, is nicely worded in Black et al. (2005a):

> "If the condition states correspond to intervals of an underlying continuous condition measure, then an item is likely to enter the state near one boundary and then progress over the time periods to near the other boundary before leaving the state. Hence the transition probability is likely to increase as the item approaches the second boundary, and so the transition probability could often increase with the time spent in the state."

As most classification schemes do not assume an underlying continuous process, like is illustrated with the quote on page 23, this reasoning is not relevant in most cases.

Summary

This chapter reviews the nature of bridge inspection data and various approaches which have been suggested and applied to fit a finite-state Markov process to this data. Most statistical models ingnore the fundamental assumption of Markov processes; namely that the future states are independent of past states given the current state. These models use the state

observations as individual and independent observations, whereas there should be dependence between successive observations of the same structure.

A better approach is to use the maximum likelihood principle, but the Poisson regression model on page 29 is too restrictive and the multinomial model on page 32 can only be used if the structure is continuously monitored. Other approaches, like a Bayesian approach or a nonparametric model, suffer from various shortcomings which make them unattractive for decision makers.

Practically all bridge condition data, obtained using visual inspections, is in the form of panel data; that is, throughout the year inspections are periodically performed at different intervals which results in a form of interval censoring. This heavily censored data can not be used to adequately test the validity of the Markov assumption. However, the Markov assumption is made by the decision maker in order to develop a tractable deterioration model. Even if the underlying deterioration process does not possess the Markov property, the assumption is necessary to be able to efficiently calculate the outcome, which is required for decision making.

3

Proposed framework

Given the conclusions drawn in the previous chapter on the various approaches to fitting finite-state Markov processes to bridge condition data, this chapter describes the appoach selected for application to bridge condition data in the Netherlands. The likelihood function plays a central role in this approach and may be defined for perfect and imperfect inspections. The calculation of the likelihood of the observed bridge conditions involves the calculation of the transition probability function $P_{ij}(s,t) = \Pr\{X(t) = j \,|\, X(s) = i\}$, which gives the probability of a transition from state i to state j between ages s and t for $s \leq t$. The calculation of $P_{ij}(s,t)$ is not straightforward, therefore several methods are presented separately in Chapter 7. Attention is also given to covariate analysis, which tests if considering different groups of bridges will considerably influence the outcome of the model. Also, Section 3.4 lists the data requirements for the successful application of this estimation procedure.

3.1 MAXIMUM LIKELIHOOD ESTIMATION

For a given set of data, the likelihood is defined as the probability that the chosen model generates the data. It is simply the joint density of the random variables of the model. For the purpose of estimation, the likelihood is considered to be a function of the unknown model parameters (Mood et al., 1974, p.278).

Given the sample set $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, the likelihood that the set of parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$ generates the data is given by the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x})$. The notation $(\boldsymbol{\theta}; \boldsymbol{x})$ means that it is a function of $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$, but that it is not a conditional probability. Obviously, $0 \leq L(\boldsymbol{\theta}; \boldsymbol{x}) \leq 1$. The problem of estimation is now reduced to a problem of maximizing the likelihood function; that is, to find those values for the model parameters which maximize the likelihood function. An equivalent problem arises when taking the natural logarithm of the likelihood function. The locations of the maxima of the log-likelihood function $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log L(\boldsymbol{\theta}; \boldsymbol{x})$ and the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x})$ are the same. It is generally easier to work with the logarithm of the likehood function as opposed to working with the likelihood function itself. This is especially true if the observations of the model are considered to be independent. Let

$f_M(x)$ be the probability of taking a sample x from the model M, then under the assumption of independent sampling the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log \prod_i f_M(x_i) = \sum_i \log f_M(x_i) \tag{3.1}$$

for samples $x_i, i = 1, 2, \ldots$. Here, the dependence of the model M on the set of parameters $\boldsymbol{\theta}$ is surpressed; i.e., $M \equiv M(\boldsymbol{\theta})$.

The use of maximum likelihood estimation for panel data is quite common. The likelihood function as given by Equation (3.1) can readily be found in the literature. For example, it is common in publications in the field of medical statistics: Jackson et al. (2003) provide an overview, which includes Kay (1986). The latter discusses the application of a continuous-time Markov process to the development of cancer markers over time. These markers are used to grade levels of disease in order to identify various states in the development of an illness. An important reference is the paper by Kalbfleisch and Lawless (1985), which includes many of the topics discussed in this chapter and uses an example of a longitudinal study of smoking habits of school children. Like many studies, the assessment of smoking habits is done at infrequent moments in time at which a large group of subjects is evaluated at the same time. This results in panel data and Kalbfleisch and Lawless (1985) used Fisher's method of scoring in combination with the diagonalization method for calculating the transition probability function (see page 115 in Chapter 7). The method of scoring will be presented in Section 3.3.

Other than the fact that it is particularly well suited for application to panel data, maximum likelihood estimation has the additional benefit that, under suitable conditions and if the sample set is sufficiently large, the estimator is approximately Gaussian (i.e., normally distributed). According to a well known theorem (see e.g., Mood et al. (1974, Theorem 18 on page 359)) the estimator $\hat{\boldsymbol{\theta}}_*$ of the true value $\boldsymbol{\theta}_*$ asymptotically has an m-dimensional multivariate normal distribution: $\hat{\boldsymbol{\theta}}_* \sim \mathrm{MVN}_m(\boldsymbol{\theta}_*, I^e(\boldsymbol{\theta}_*)^{-1})/n$. Here, $I^e(\boldsymbol{\theta})$ is the expected information matrix, which will be discussed in the next section.

The concept of maximum likelihood estimation was developed in the early 20th century by British statistician R.A. Fisher (1912 and 1922) and is now a common method for statistical inference. In the following section, the likelihood functions for both perfect and imperfect periodic observations of a finite-state Markov process are derived.

3.2 STATISTICAL MODEL

As the model under consideration is a stochastic process, an observation of the model M is a collection or sequence of one or more state observations at successive times $t_1 < t_2 < \cdots < t_n$. The process starts at $t_0 = 0$ and it is assumed here that the initial state $X(t_0)$ is known to the decision maker. For ease of notation, $X_k \equiv X(t_k)$ is used in the following discussion.

3.2.1 LIKELIHOOD FUNCTION FOR PERFECT OBSERVATIONS

The likelihood function for perfect observations is quite straightforward. With 'perfect', it is meant that no error is made in the assessment of the true condition state of the structure or component. For each object, a sequence $\{X_1, X_2, \ldots, X_n\}$ of state observations is available. The probability of each sequence may conveniently be written as a product of transition probability functions:

$$\Pr\{X_1, X_2, \ldots, X_n\} = \Pr\{X_0\} \prod_{k=0}^{n} \Pr\{X_{k+1} \,|\, X_k\}, \tag{3.2}$$

where the probability distribution of the initial condition state $\Pr\{X_0\}$ is assumed to be known.

Special attention should be given to the case where an observed transition is not possible given the model. In this case, $\Pr\{X_{k+1} \,|\, X_k\} = 0$ and taking the logarithm will result in an error. This situation can arise if only deterioration is allowed, but an improvement in the condition is observed. These cases should simply be discarded by the implementation. There are two options for how to deal with these cases: discard the impossible transition or discard the complete sequence of observations of which the transition is a part. In order to use as much data as possible, the first option is preferred. However, this may be subject to debate because it may seem more natural to choose for the second option. In the next section, a model which allows for imperfect inspections will be presented and the algorithm used in this model discards the complete sequence of observed bridge conditions.

3.2.2 LIKELIHOOD FUNCTION FOR IMPERFECT OBSERVATIONS

In order to model the subjectiveness of visual inspections and the natural variability which arises from this, it is possible to consider that inspections may not accurately describe the true state of the process. By 'imperfect' it is meant that there exists a true state and that, with some probability, an inspection may indicate a different state other than the true state. Other suitable designations for this topic include 'inspection accuracy' or

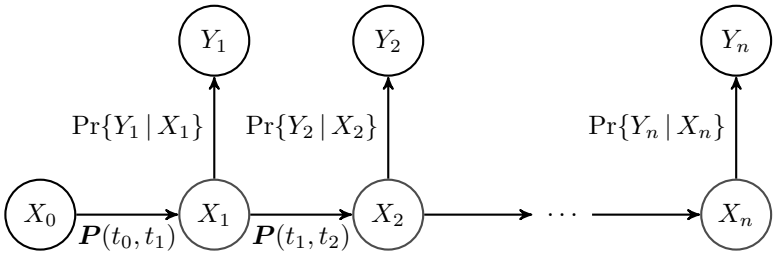


Figure 3.1: Graphical representation of a hidden Markov model.

'inspection variability'. It is noted that an inspector can never be blamed for making 'inaccurate' observations, as he or she is required to give a personal assessment which, in his or her view, best reflects the condition state.

This subject has been of great interest in the past and with good reason. It is obvious that the personal interpretation of condition states, like those in Tables 1.1 and 2.1, and how these relate to damages on structures will vary substantially between inspectors. A study by Phares et al. (2002) has shown that there is indeed a substantial difference between the ratings from different inspectors. This variability may also account for some of the increases in the quality of structures which can not be attributed to maintenance. In general, data sets from inspections will include transitions towards better conditions, which can not be eliminated on the basis of being a result of maintenance. Since a monotonically increasing Markov process for modeling deterioration can not cope with condition improvements, these pose a challenge for the decision maker. In various studies, such as those conducted by Kallen and van Noortwijk (2005b) and Morcous (2006), the condition improvements are simply removed from the data.

A modeling approach for incorporating imperfect inspections and which is commonly applied to problems in speech recognition, is the use of a hidden Markov model. Although visual inspections are the only means by which the state of deterioration may be measured, the hidden Markov approach assumes that there is some true condition state which is 'hidden' to the observer. Figure 3.1 shows a graphical representation of the hidden Markov model.

In this figure, the true states $X_1, X_2, \ldots, X_n$ at the inspection times $t_1, t_2, \ldots, t_n$ are hidden behind the actual observed states $Y_1, Y_2, \ldots, Y_n$. The probability of the observations, given the true state, is given by $e_{ij} = \Pr\{Y_k = j \mid X_k = i\}$, where $0 \leq e_{ij} \leq 1$ and $\sum_j e_{ij} = 1$. The (discrete) probability distribution $\boldsymbol{e}_i = \{e_{i1}, \ldots, e_{in}\}$ reflects the variability in the inspections. These 'error' distributions can be assessed by expert judgment, testing a pool of experts using selected test cases, or by maximum

likelihood. The decision maker must also choose how large the inspectors' mistakes are allowed to be. He may restrict the model such that only an error of one state (higher or lower) can be made.

The observed sequence $Y_1, Y_2, \ldots, Y_n$ has several properties, but it does not possess the Markov property. One of the primary problems involved with the use of hidden Markov models, is determining the probability of the observed sequence given the model. This probability is required for maximum likelihood estimation. Most properties rely on the basic property that, conditional on the true process, the observations Y_k for $k = 1, \ldots, n$, are independent; that is:

$$\Pr\{Y_1, \ldots, Y_n \,|\, X_1, \ldots X_n\} = \prod_{k=1}^{n} \Pr\{Y_k \,|\, X_k\}.$$

The probability of the observed sequence may be determined by

$$\Pr\{\boldsymbol{Y}\} = \sum_{\forall \boldsymbol{X}} \Pr\{\boldsymbol{Y}, \boldsymbol{X}\} = \sum_{\forall \boldsymbol{X}} \Pr\{\boldsymbol{Y} \,|\, \boldsymbol{X}\} \Pr\{\boldsymbol{X}\}, \tag{3.3}$$

where $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ and $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. The summation in Equation (3.3) is taken over all possible sequences $\boldsymbol{X}$. As this approach is computationally very inefficient, there exists the 'forward-backward' algorithm. The forward part of this algorithm is most frequently used to calculate the likelihood of the observed sequence. Let the forward variable be defined as

$$\alpha_i^{(k)} = \Pr\{Y_1, Y_2, \ldots, Y_k, X_k = i\},$$

which is the probability of the observations up to time t_k and the true state at time t_k being equal to i. The forward algorithm starts with $k = 1$:

$$\alpha_i^{(1)} = \Pr\{Y_1, X_1 = i \,|\, X_0\} = \Pr\{Y_1 \,|\, X_1 = i\} \Pr\{X_1 = i \,|\, X_0\}$$

and continues for $k > 1$ using the relation

$$\alpha_i^{(k+1)} = \Pr\{Y_{k+1} \,|\, X_{k+1} = i\} \sum_j \alpha_j^{(k)} \Pr\{X_{k+1} = i \,|\, X_k = j\} \tag{3.4}$$

up to $\alpha_i^{(n)}$. The likelihood of the complete sequence of observations may then be obtained by summing over all X_n:

$$\Pr\{Y_1, \ldots, Y_n\} = \sum_i \alpha_i^{(n)}.$$

This approach is also referred to as the Baum-Welch algorithm. The properties of hidden Markov models, on which this algorithm relies, are discussed and proven in MacDonald and Zucchini (1997).

Note that this algorithm will set the likelihood of the complete condition history of a structure to zero if one of the transitions is not possible with the given model. This can be observed in Equation (3.4):

$$\text{if } \alpha_i^{(k)} = 0 \text{ for all } i, \text{ then } \sum_i \alpha_i^{(n)} = 0 \text{ for } n \geq k.$$

In order to check the correctness of the implementation of this model, one can take the probability of a misclassification to be zero and compare the results of the model with those obtained from the model with perfect inspections. For this, one should set $e_{ii} = 1$ and $e_{ij} = 0$ for $i \neq j$ and let the model for perfect inspections discard complete observation sequences if one of the transitions in the sequence is not possible. Also, the probability of the first observed condition must be one, i.e. $\Pr\{Y_1 = y_1\} = 1$ for the observed state y_1.

Some other approaches to this problem can be found in the literature. For example, Jackson et al. (2003) formulate Equation (3.3) as a product of matrices, which is just the forward-backward algorithm in matrix formulation; see also MacDonald and Zucchini (1997, p.61). In Cappé et al. (1998), the likelihood of the observed sequence is rewritten as

$$\begin{aligned}
\Pr\{Y_1, \ldots, Y_n\} &= \prod_{k=1}^{n} \Pr\{Y_k \,|\, Y_1, \ldots, Y_{k-1}\} \\
&= \prod_{k=1}^{n} \sum_j \Pr\{Y_k, X_k = j \,|\, Y_1, \ldots, Y_{k-1}\} \\
&= \prod_{k=1}^{n} \sum_j \Pr\{Y_k \,|\, X_k = j\} \Pr\{X_k = j \,|\, Y_1, \ldots, Y_{k-1}\}.
\end{aligned}$$

Using the logarithm of this result, they propose a recursive algorithm which also enables the calculation of the derivatives with respect to the model parameters iteratively.

The name 'hidden' is primarily used in the area of speech recognition, where Rabiner (1989) is often cited. In the field of operations research, the same model is referred to as a 'partially observable' Markov model. The latter has been applied to the problem of bridge management by Ellis et al. (1995), Jiang et al. (2000) and Corotis et al. (2005). The same model has also been referred to as a 'latent' Markov model and applied to bridge management by Madanat (1993) and Smilowitz and Madanat (2000). In all of these publications, the use of a hidden Markov model is combined with a Markov decision process (see the section on Markov decision processes on page 87).

3.2.3 COVARIATE ANALYSIS

Covariate analysis is often used in the field of survival analysis and aims to determine which 'covariates' influence the outcome of the model. Covariates are simply additional variables, also called 'independent' variables, which take on two or more values depending on the properties of a structure. In medical applications, a very common covariate is the sex of patients. As an example, let X_s represent this covariate and let $X_s = 0$ if the patient is a male and $X_s = 1$ if the patient is a female. In essence, covariate analysis groups the objects according to one of their properties and then assesses the impact of this grouping on the estimate of the parameters. If, in the example of the patients, the outcome is significantly different, it must be concluded that a patient's sex is a statistically significant variable in the model. Using a mathematical formulation: suppose one wants to estimate the transition intensity $\lambda > 0$ in a continuous-time Markov processes, then the covariate model looks like

$$\lambda_s = \exp\{\beta_0 + \beta_1 X_s\}, \tag{3.5}$$

where β_0 is the intercept and β_1 is the coefficient of the variable X_s. Using this reparametrization, the 'new' parameters $\boldsymbol{\beta} = \{\beta_0, \beta_1\}$ are estimated. The model in Equation (3.5) is referred to as a multiplicative model, where $\exp\{\beta_0\}$ represents the 'base rate' and $\exp\{\beta_1 X_s\}$ is the adjustment due to X_s. If β_1 turns out to be close to zero, such that the adjustment is close to unity, it may be concluded that grouping of patients according to their sex does not influence the outcome of the estimation. A formal test like a likelihood ratio hypothesis test may be used to validate or reject the hypothesis that a covariate is not statistically significant.

In Skuriat-Olechnowska (2005), a simple Wald test was used for the purpose of testing the statistical influence of each covariate in the analysis. Using the fact that the maximum likelihood estimator for each coefficient β_k, $k = 0, 1, 2, \ldots$, is asymptotically normal, the Wald test defines the Z-statistic as the ratio

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}_{\hat{\beta}_k}},$$

where $\mathrm{SE}_{\hat{\beta}_k}$ is the standard error of the estimator $\hat{\beta}_k$. The standard error is the estimated standard deviation of the estimator. The null hypothesis is that the covariate i does not significantly influence the transition intensities. In other words: that the coefficient β_k is not significantly different from zero. The hypothesis $\beta_k = 0$ is rejected if the two-sided p-value $\Pr\{|Z_k| > \hat{\beta}_k/\mathrm{SE}_{\hat{\beta}_k}\}$ is small. Basically, the Wald test rejects the hypothesis at (for

example) the 5% significance level if zero is not inside the 95% confidence bounds of the estimated coefficient β_k.

The log-linear model $\lambda = \exp\{\boldsymbol{\beta X}\}$, with $\boldsymbol{X} = \{1, X_1, X_2, \ldots, X_n\}$ is a convenient model, because the positive exponential function ensures that the requirement $\lambda > 0$ always holds. For discrete-time Markov processes, the transition probabilities p may be reparametrized as

$$p = \left(1 + \exp\{-\boldsymbol{\beta X}\}\right)^{-1}, \tag{3.6}$$

such that $0 \leq p \leq 1$. Both approaches are particularly well suited for use in a maximization procedure, which is the topic of the next section.

3.3 MAXIMIZATION

In most practical cases, it is infeasible to analytically maximize the log-likelihood function, therefore a numerical approach is used. Concepts like Newton's method, quasi-Newton methods, and Fisher's method of scoring are introduced in this section. Fisher's method of scoring is a quasi-Newton method specifically used for maximizing likelihood functions and quasi-Newton methods are approximations to Newton's method, which is an iterative approximation scheme with reasonably fast (quadratic) convergence to the root of a differentiable function.

Newton's method

Newton's method (also referred to as Newton-Raphson) is a root-finding algorithm for (systems of) non-linear and continuously differentiable functions. For a vector $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ and $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function, the algorithm attempts to iteratively approximate a root $\boldsymbol{x}_*$ for which $f(\boldsymbol{x}_*) = 0$. The method is based on a Taylor series expansion of the function $f(\boldsymbol{x})$ around a value $\boldsymbol{a}$: $f(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{a})f'(\boldsymbol{x})$, or if solved towards $\boldsymbol{a}$: $\boldsymbol{a} = \boldsymbol{x} - f(\boldsymbol{x})/f'(\boldsymbol{x})$. If $\boldsymbol{a} = \boldsymbol{x}$, then $f(\boldsymbol{x}) = 0$ and the aim of Newton's method is to approximate the unknown $\boldsymbol{x}_*$ with iterative adjustments to $\boldsymbol{a}$. This is done by taking an initial guess $\boldsymbol{x}_0$ for $\boldsymbol{x}_*$ and successively calculating $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ using the relation

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{f(\boldsymbol{x}_k)}{f'(\boldsymbol{x}_k)}, \tag{3.7}$$

for $k = 0, 1, 2, \ldots$ until $\| \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \| < \epsilon$. This algorithm is not guaranteed to converge, but it converges fairly fast when it does converge. If there are multiple (local) extremes in the function f, then the algorithm is also not guaranteed to converge to the global extreme.

The algorithm can also be used for a system of non-linear equations. Let $\boldsymbol{f}(\boldsymbol{x}) = \{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\}$, then Equation (3.7) becomes

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{J}^{-1}(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}_k), \tag{3.8}$$

where $\boldsymbol{J}^{-1}(\boldsymbol{x})$ is the inverse Jacobian of $\boldsymbol{f}(\boldsymbol{x})$. For the system of equations $\boldsymbol{f}(\boldsymbol{x})$, the Jacobian is the $n \times m$ matrix defined as

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

The location of the extremes of a function $f(\boldsymbol{x})$ may be determined by setting the partial derivatives to each of the variables x_i equal to zero and solving for x_i. For $\boldsymbol{x} \in \mathbb{R}^n$, this results in the requirement to solve a system of n (non-linear) equations:

$$\frac{\partial f}{\partial x_1} = 0, \frac{\partial f}{\partial x_2} = 0, \ldots, \frac{\partial f}{\partial x_n} = 0.$$

Let

$$\boldsymbol{f}'(\boldsymbol{x}) = \left\{ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right\},$$

then Newton's step for iteratively approximating $\boldsymbol{x}_*$, such that $f(\boldsymbol{x}_*) = 0$, becomes

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{H}^{-1}(\boldsymbol{x}_k)\boldsymbol{f}'(\boldsymbol{x}_k) \tag{3.9}$$

for $k = 0, 1, 2, \ldots$. Here, $\boldsymbol{H}^{-1}(\boldsymbol{x})$ is the inverse Hessian matrix of $f(\boldsymbol{x})$ defined as

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

If $f(\boldsymbol{x})$ is a log-likelihood function, then $f'(\boldsymbol{x})$ and $-\boldsymbol{H}(\boldsymbol{x})$ are referred to as the score function (or efficient score) and the information respectively. In this context, define

$$s_{\theta_i} = \frac{\partial}{\partial \theta_i} \ell(\boldsymbol{\theta}; \boldsymbol{x}),$$

and in vector notation:

$$\boldsymbol{s} = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{x})$$

as the score function. The information is denoted by $\boldsymbol{I}^o(\boldsymbol{\theta}; \boldsymbol{x}) = -\partial \boldsymbol{s} / \partial \boldsymbol{\theta} = -\boldsymbol{H}(\boldsymbol{x})$. The dependence of $\boldsymbol{s}$ and $\boldsymbol{H}$ on $\boldsymbol{\theta}$ is suppressed in the notation.

Note that Newton's method is a method for unconstrained optimization, which means that the parameters of the function to be minimized are not constrained. As most practical applications have some restrictions on the domain of the parameters, a reparametrization is used. For continuous-time or discrete-time Markov processes, Equations (3.5) and (3.6) may be used respectively.

Quasi-Newton methods and Fisher's method of scoring

The adjustment in Equation (3.7) and Equation (3.8) contains the first derivative of the function f with respect to each parameter and the Newton step in Equation (3.9) also requires the second derivatives. In most cases, these derivatives are not easily obtained, therefore they are replaced by approximations which do not require the derivatives themselves. When using this approach, Newton's method is commonly referred to as a quasi-Newton method.

Fisher proposed to replace the information with the expected value (over all possible realizations of the model) denoted by $\boldsymbol{I}^e(\boldsymbol{\theta}; \boldsymbol{x}) = \mathbb{E}[\boldsymbol{I}]$. This approach eliminates the need to calculate the second derivatives of the likelihood function. This can be shown as follows: let $f(\boldsymbol{x}; \boldsymbol{\theta}) = L(\boldsymbol{\theta}; \boldsymbol{x})$ and $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log f(\boldsymbol{x}; \boldsymbol{\theta})$, then

$$\begin{aligned} I_{\theta_i \theta_j} &= -\frac{\partial}{\partial \theta_j} \left[\frac{1}{f} \frac{\partial f}{\partial \theta_i} \right] \\ &= -\left[-\frac{1}{f^2} \frac{\partial f}{\partial \theta_j} \frac{\partial f}{\partial \theta_i} + \frac{1}{f} \frac{\partial^2 f}{\partial \theta_j \partial \theta_i} \right] \\ &= s_{\theta_j} s_{\theta_i} - \frac{1}{f} \frac{\partial^2 f}{\partial \theta_j \partial \theta_i}. \end{aligned}$$

Multiplying each side with the density $f(\boldsymbol{x}; \boldsymbol{\theta})$ and integrating over all possible values of $\boldsymbol{x}$ results in

$$\int_{\boldsymbol{x}} I^o_{\theta_i \theta_j} f \mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{x}} s_{\theta_j} s_{\theta_i} f \mathrm{d}\boldsymbol{x} - \int_{\boldsymbol{x}} \frac{\partial^2 f}{\partial \theta_j \partial \theta_i} \mathrm{d}\boldsymbol{x}.$$

As

$$\int_{\boldsymbol{x}} \frac{\partial^2 f}{\partial \theta_j \partial \theta_i} = \frac{\partial^2}{\partial \theta_j \partial \theta_i} \int_{\boldsymbol{x}} f \mathrm{d}\boldsymbol{x} = 0,$$

this reduces to $\mathbb{E}\left[\boldsymbol{I}_{\theta_i\theta_j}\right] = \mathbb{E}\left[s_{\theta_j} s_{\theta_i}\right]$. Because it is generally not feasible to calculate the expected information, the 'observed' information $\boldsymbol{I}^o_{\theta_i\theta_j}$ is determined instead. The observed information is approximation of the expectation by the mean over the observed data. For multiple (independent) observations of the model, the score and information become

$$s_{\theta_i}(\boldsymbol{\theta}; \boldsymbol{x}) = \frac{\partial}{\partial \theta_i} \log \prod_{\forall k} f_{M(\boldsymbol{\theta})}(x_k) = \sum_{\forall k} \frac{\partial}{\partial \theta_i} \log f_{M(\boldsymbol{\theta})}(x_k) = \sum_{\forall k} s_{\theta_i}(\boldsymbol{\theta}; x_k)$$

and

$$\begin{aligned} \boldsymbol{I}^o(\boldsymbol{\theta}; \boldsymbol{x}) &= -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \prod_{\forall k} f_M(x_k; \boldsymbol{\theta}) = \sum_{\forall k} -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_M(x_k; \boldsymbol{\theta}) \\ &= \sum_{\forall k} \boldsymbol{I}^o(\boldsymbol{\theta}; x_k) \end{aligned}$$

respectively. Therefore, the total score and information are simply the sum of the individual scores and informations.

3.4 Data requirements for model application

There are several requirements which have to be met if bridge inspection data is to be used in a maximum likelihood estimation procedure as described in this chapter. First, there have to be a minimum number of inspections available for each structure if it is to contribute to the likelihood function. For a model with perfect inspections, these should be

- at least two successive state observations if the bridge was constructed before registration of inspection results was initiated,
- and at least one if the bridge was constructed after registration of inspection results was initiated.

If inspections are assumed to be imperfect, the requirements are that

- there is at least one state observation available if the bridge was constructed before registration of inspection results was initiated and that an assumption be made about the actual state distribution at the time of the first inspection,
- and there is at least one state observation available if the bridge was constructed after registration of inspection results was initiated and the initial state distribution at service start of the bridge is given.

The assumption about the distribution of the actual state at the first inspection, is required due to the fact that there is no information about the

state at the service start of the structure or on past maintenance activities. The distribution of X_1 conditional on X_0 is therefore not available and should be replaced by a distribution over X_1 selected by the decision maker. There is no exact rule for the minimum number of state histories (i.e., structures available for the analysis) which must be available for a successful application of maximum likelihood estimation. A simple guideline is: the more, the better.

SUMMARY

This chapter presents the theoretical foundation for the maximum likelihood estimation of transition intensities in a continuous-time Markov process. The maximum likelihood approach is particularly well suited for panel data, because it considers the likelihood of the periodic state observations and it uses the transition probabilities to determine this likelihood. The method therefore respects the Markov assumption and it does not rely on information which is not available, namely information on the exact times of transitions or the exact length of stay in any state. For the application of this method, the data must fulfill certain requirements, which are outlined in the previous section.

The likelihood function can be defined to achieve several goals: estimate the model parameters under the assumption that state observations are perfect (Section 3.2.1) or imperfect (Section 3.2.2), or to determine the influence of dependent variables in a covariate analysis (Section 3.2.3). The likelihood must then be maximized as a function of the model parameters, such that the parameter values are the most likely ones to have generated the given data. The most common method for maximizing the likelihood function is Fisher's method of scoring, which is discussed in Section 3.3.

4

Application and results

In this chapter, the approach proposed in Chapter 3 will be applied to bridge condition data obtained in the Netherlands. A general introduction to the practice of inspecting bridges in the Netherlands was given in Section 1.3 on page 6. The most important aspect of bridge inspections is that individual damages are registered in the database and that the severity of these damages is then used to set the condition of the structure. The severity of the damages is also used to set the condition of the basic and primary components on which the damages are located.

First, a description of the available data sets will be given. Then, various models with different transition structures are fitted to the data sets. In Section 4.2.1, this is done for progressive Markov processes which only allow for deterioration, and in Section 4.2.2, this is done for models which also allow for transitions to better states. Sections 4.3 and 4.4 respectively discuss the application of hidden Markov models and the analysis of covariate influences on the model parameters.

4.1 DUTCH BRIDGE CONDITION DATA

In the Netherlands, bridge inspections register the location and severity of all damages present on the structure. These severities are then used by the database and by the inspector to set the condition of the individual components and the overall condition of the structure. In order to succesfully fit a deterioration model, it is necessary to have a sufficient amount of data. This is especially true for fitting stocastic processes for modeling uncertain deterioration over time.

4.1.1 QUALITY AND DETAIL OF INFORMATION

Although it is not mandatory, the cause of each damage may be registered in the database. From a decision maker's point of view, it may be interesting to identify different physical deterioration mechanisms and to fit the deterioration model to the historic development of the damages. However, it is not possible to identify the same damages between successive inspections in the database. This is because the identification of the location of each damage is done by issuing a location number on a technical drawing

of the structure. Inspectors issue these locations anew at each inspection, therefore the same damage may have a different location number at different inspections. From a practical point of view, it is not feasible to visually check the location of damages for all inspections in the database. After filtering for incorrect or incomplete entries, there are roughly 6000 inspection events on 2300 structures which can be used for modeling deterioration with Markov processes.

There are various reasons why some inspection events or structures may not be included in the data sets.

- A structure may not have a construction year assigned to it, which is necessary information for determining the age of structures at the time of an inspection. The construction year may also be faulty; for example, it may be after 2004 which is the last year in which inspections were registered in the database.
- An inspection event may have an incorrect date assigned to it, where the year of the inspection is before the construction year of the structure or after 2004.
- A structure may not have undergone enough inspections. In general, structures should at least have been inspected twice, such that two reference points in time are available. For structures built after 1985 only one inspection is required, because the initial state (i.e., perfect or state '0', see Table 1.1) can be used as the first 'observation'. For bridges constructed before 1985, it is not known if maintenance was performed before the registration in the database began.

It is a not an easy task to extract the data from the relational database and putting this information in a form suitable for estimating the model parameters. Aside from the easily identifiable faulty data mentioned above, there are an unknown number of entries in the database which are also faulty, but which can not easily be identified as such. One such example occurs when an inspector incorrectly registers a condition 0, when the purpose of the inspection was not to assess the condition of the structure. Some inspections are performed to investigate specific areas or damages on the structure. These inspections should be registered, but no condition number should be entered into the database. The analysis of textual remarks in the database may be able to filter out these entries, but this requires a significant amount of work.

Another area where mistakes can enter the database is in the dates of inspections. The date which is registered in the database may not actually be the date at which the inspection physically took place. The durations between the inspection dates in the database are collected in a histogram in Figure 4.1. Even though inspections are performed periodically, there's quite a bit of variability in the actual times between these inspections. This

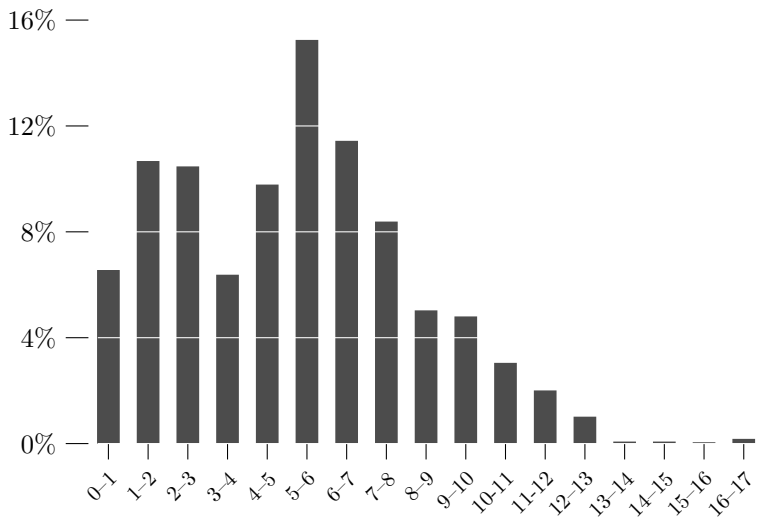


Figure 4.1: Histogram of the number of years between bridge inspections in the Netherlands.

is mainly due to the actual planning and execution of inspections resulting in different inspection intervals. Some of the differences will also be due to the fact that the date in the database does not always correspond to the actual date of the inspection.

All the anomalies mentioned up to now are the result of inconsistent or erroneous behaviour by the inspectors. There is however another type of faulty data: the absence of information on when maintenance was performed on the structure. When faced with an increase in the condition of a structure, the decision maker would like to attribute this increase to either maintenance or to a difference of opinion between two inspectors. Although the condition database in the Netherlands has the capability to register maintenance activities, this feature has seldomly been used. Morcous (2006) reports the same lack of maintenance data in the bridge condition data for Québec in Canada.

From this discussion, one may conclude that the quality of the data is insufficient for use in maintenance management. However, some of the inconsistencies are the inevitable result of the long duration of 20 years of operating the database and of the many people involved in the inspections. Even after filtering for faulty data, the number of observed bridge conditions is very large and is certainly suitable for use in a statistical analysis.

As mentioned in Section 1.3 of the introduction, the inspections do not lead to a uniform observation of all condition states. The histogram in

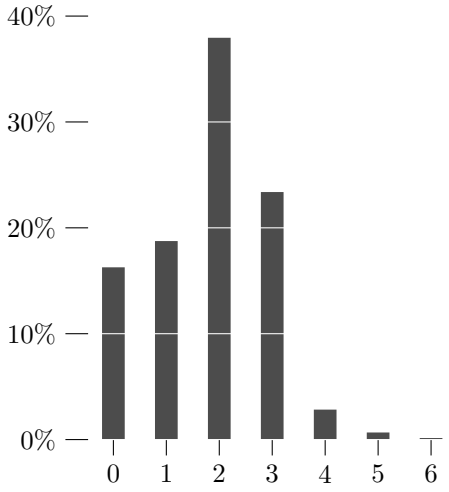


Figure 4.2: Histogram of the observed states in the bridge condition database in the Netherlands.

Figure 4.2 shows that states 4, 5 and 6 of the condition scale in Table 1.1 are observed much less often than states 0 to 3. Because the last state, namely state 6, is extremely rare, it has been combined with state 5 to form a new state representing the conditions 'bad' and worse. The number of states considered in this chapter, is therefore equal to six.

4.1.2 Extraction from database

The original database is a relational database, which means that the information is stored in various tables which are related to eachother. The tables containing the basic information on the structures (e.g., name, location, size, etc.) are relatively static as changes in this information are not frequent. The information obtained through inspections constitutes the more dynamic part of the database. For each inspection, an entry is made in a table and each entry is linked to the relevant structure in another table. This allows information to be added efficiently and in a structured manner.

For the purpose of a statistical analysis, the relational database is not the most convenient form for data storage. Therefore the data is extracted from the database and stored in a spreadsheet. For each inspection, a new line containing all necessary information for the analysis is written to the spreadsheet. This includes at least a unique identifier, the year of construction, the inspection date and the condition state. Other information, for

example for use in a covariate analysis, may also be included. The final step in the extraction process, is to prepare the data for use in a program for numerical analysis. For maximum portability, the data is written to a plain text file. There are two options for this: 1) each line contains a transition, or 2) each line contains only an observation. The first option results in a file of the form

```
1       249     341     1       3
2       213     306     1       3
3       213     309     1       2
...
17      248     261     1       2
17      261     351     2       3
...
```

Here, the first column includes a unique identifier. The second and third columns contain the age in months at two consecutive inspections and the fourth and last columns contain the condition states at these inspections. The second option contains the same data in a different form:

```
1       249     1
1       341     3
2       213     1
2       306     3
3       213     1
3       309     2
...
17      248     1
17      261     2
17      351     3
...
```

Each observation is on a different line, where the first column includes the identifier, the second the age in months at the inspection, and the third the observed state.

The first option can be used for estimating the parameters of the Markov process when inspections are assumed to be perfect. The second option is particularly well suited for applications where inspections may vary, but can equally well be used for models with perfect inspections.

4.1.3 DESCRIPTION OF DATA SETS

In this chapter, four data sets will be used: general condition of structures, most severe damage condition of structure, superstructure conditions, and the condition of kerbs. Each of these are briefly explained here.

General condition of structures

The general or overall condition of the structure is the condition state which is assigned to the 'logical structure' in the database as shown in Figure 1.5. This data set, and the one described next, contains the most observations of all data sets. As mentioned earlier, the condition data assigned to a structure and its components is derived from the severity of damages registered at each inspection. Therefore, the components will not be assigned a condition state at each inspection, unless there are one or more damages present on the component.

Most severe damage on structures

The general condition of a structure is automatically assigned the most severe condition of the damages present at the time of an inspection. The inspector must manually change this condition if it is not representative for the structure as a whole. As a sort of 'worst case scenario', the analysis is also performed on a data set which includes the condition indicators of the most severe damages registered at each inspection.

Condition of superstructures

The superstructure of a bridge is generally considered to be all structure above the bearings. This includes the beams, road surface, kerbs, and safety barriers. In Figure 1.5, superstructures belong to the group of principal components. This data set includes roughly 5500 observations.

Condition of kerbs

Kerbs (or curbs) are the rims along the roadway, forming an edge for a sidewalk or for the safety barriers. There are roughly 5500 observations of kerb condition states.

4.2 SELECTION OF TRANSITION STRUCTURE

Finite-state Markov processes are completely shaped by the transition structure defined in the transition intensity matrix. This matrix determines when and to where a transition can take place. Before embarking on fitting a Markov deterioration process to observed condition data, the decision maker must decide how deterioration may procede. It is possible to choose a fully filled transition matrix, which allows for transitions between any two condition states including transitions to better states. This, however, will generally not be a very useful approach. In this section, various 'models' will be fitted to the Dutch bridge condition data. These models are distinguished by the absence (without maintenance) or presence (with maintenance) of backward transitions.

4.2.1 Models without maintenance

This section discusses the estimation of transition intensities in sequential Markov deterioration processes. Structures are therefore only allowed to move to the next state and not move backwards to better states. The basic transition intensity matrix is then given by

$$\boldsymbol{Q}(t) = \begin{bmatrix} -\lambda_0(t) & \lambda_0(t) & 0 & 0 & 0 & 0 \\ 0 & -\lambda_1(t) & \lambda_1(t) & 0 & 0 & 0 \\ 0 & 0 & -\lambda_2(t) & \lambda_2(t) & 0 & 0 \\ 0 & 0 & 0 & -\lambda_3(t) & \lambda_3(t) & 0 \\ 0 & 0 & 0 & 0 & -\lambda_4(t) & \lambda_4(t) \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{4.1}$$

An unlimited number of choices are available for the elements $\lambda_i(t)$, $i = 0, \ldots, 4$, but only five options will be considered here. These options are defined in Table 4.1.

Model	$\lambda_i(t)$	n	Description
A	a	1	state- and age-independent
B	abt^{b-1}	2	state-independent and age-dependent
C	a_i	5	state-dependent and age-independent
D	$a_i bt^{b-1}$	6	state- and age-dependent
E	$a_i b_i t^{b_i-1}$	10	state- and age-dependent

where a, $a_i, t > 0$ and $-\infty < b, b_i < \infty$ for $i = 0, \ldots, 4$

Table 4.1: Five options for the transition intensity parameters in a sequential Markov deterioration process.

This table names the models A to E according to the number of parameters n included in the model. Model A has only one parameter, whereas model E has ten parameters. These roughly correspond to the models presented in Kallen and van Noortwijk (2006b), only model D has been renamed to E and an extra model with six parameters has been introduced as model D. Both models are state- and age-dependent, but compared to model E, the dependence on age is not state dependent as in model D. For model D, the age-dependent transition intensity matrix may be written as $\boldsymbol{Q}(t) = \boldsymbol{Q}f(t)$, where $f(t) = t^{b-1}$. In other words, the intensity matrix may be decomposed in an age-constant matrix and a scalar function which changes the intensity of the transitions over the age of structures if $b \neq 1$. The choice for the transition intensity function is arbitrary and $\lambda(t) = abt^{b-1}$ is chosen here to obtain a power law function for the integrated intensity function: $\Lambda(t) = at^b$.

Using the maximum likelihood framework as proposed in Chapter 3, the five models in Table 4.1 will be fitted to the available data sets. Note that these are nested models, with model E being the most general formulation, such that their relative quality of fit may be compared using standard statistical methods.

Overall bridge conditions

First, all models in Table 4.1 will be fitted to the observed overall condition states for the structures. Then the relative quality of fit will be compared such that a verdict can be made as to which model is most suitable for the data at hand. Note that the estimates for the parameters and the corresponding log-likelihood values are similar, but different from those reported in Kallen and van Noortwijk (2006b). This is because the data set used to obtain the results in the paper contained quite a few duplicate entries.

i	Model A		Model B		Model C		Model D		Model E	
	a_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i
0	0.18	1	1.44	0.66	0.61	1	0.48	1.06	1.19	0.85
1	0.18	1	1.44	0.66	0.40	1	0.29	1.06	0.10	1.24
2	0.18	1	1.44	0.66	0.12	1	0.08	1.06	0.10	1.04
3	0.18	1	1.44	0.66	0.04	1	0.02	1.06	0.17	0.77
4	0.18	1	1.44	0.66	0.12	1	0.08	1.06	0.32	0.86

Table 4.2: Estimated parameter values for the models without maintenance, where the unit of time is 1 year.

The results of the maximum likelihood estimation for models A to E are presented in Table 4.2. Even with an ever evolving set of data, the result for, for example, model A is close to the result reported in Kallen and van Noortwijk (2005b). The estimated annual transition intensity is 0.18, such that the mean waiting time in each state is approximately 5.5 years. The best estimate for the parameters in model B, where the transition rates also depend on the age of the structure, results in a decreasing intensity rate as the structures increase in age. Model C, with state-dependent transition rates, shows decreasing rates as the condition worsens. The same can be seen with model D, but with lower rates which increase slightly with increasing age. Model E, with state- and age-dependent transition intensities shows a mixed picture. The decreasing transition intensity in model B can be explained by the fact that bridges spend less time in initial states compared to later states, which can be seen in the parameters of model C in Table 4.2. Since there is a positive correlation between the

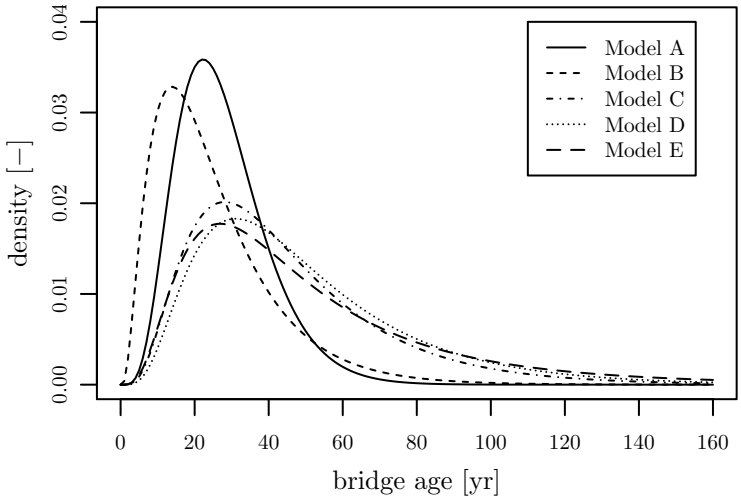


Figure 4.3: The probability density of the time to reach the final state for the models without maintenance.

age and state of a bridge, bridges have a higher transition intensity in the earlier stages of life.

Using the results in Table 4.2, it is possible to determine the time required to reach the last state, namely state 5, when starting from the initial state, namely state 0. The uncertainty in this time is represented by the 'lifetime' distributions in Figure 4.3.

Because state 5 is not considered to be a failure state, these random times do not represent true lifetime distributions. Therefore, they should not be compared to the Weibull lifetimes estimated by van Noortwijk and Klatter (2004). State 5 should be considered as an undesirable condition state and a structure in this state should be attended to within a short period of time. The mean time to reach state 5 and the 90% confidence bounds (represented by the 5% and the 90% percentiles) for the five models are listed in Table 4.3.

From this table and the results in Figure 4.3 it can be concluded that the models can be roughly grouped in two categories: models without state dependence (A and B) and models with state-dependence (C, D and E). The models without state dependence result in a much more narrow distribution compared to those with state-dependence. The difference between the times to reach state 5 for models C, D and E, which have different waiting time distributions for each state, is very small. The mean time to reach state 5 for these three models is between 45 and 50 years, which is roughly halfway the design life of 80 to 100 years for bridges in the Netherlands. If state 5 corresponds to a condition level where major renovation

Model	5% [yr]	Mean [yr]	95% [yr]
A	11	28	50
B	6	25	57
C	14	45	96
D	16	49	107
E	14	50	126

Table 4.3: Mean value and the 5% and 95% percentiles of the time (in years) to reach the final state for the models without maintenance.

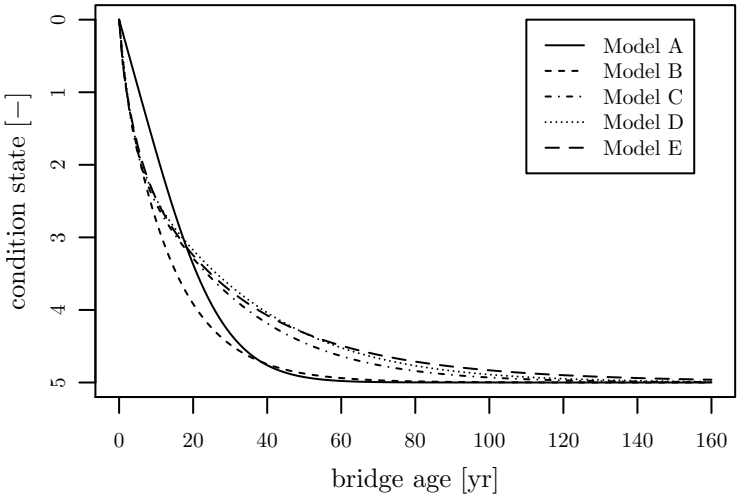


Figure 4.4: The expected condition state as a function of age for the models without maintenance.

is required, then this result is very much in line with current experience in the Netherlands. A rule of thumb in the Netherlands is that bridges require major repairs or renovation around 40 years of age. A noticable difference between models C, D and E is the increasing uncertainty towards higher ages represented by an increasingly longer tail in the distribution.

Another interesting result is the expected condition state over time as predicted by the Markov deterioration model. The same grouping according to state-dependence or -independence is observed in the expectation of the deterioration as is shown in Figure 4.4. All models converge to state 5, which is an absorbing state in all five models. The linearity of the expectation for model A in the initial states can clearly be observed.

Relative quality of fit

The likelihood value of the data given a model, is an indicator for the quality of fit relative to the likelihood values of the other models. The larger the likelihood value of the model, the larger the probability that the data could be generated by the model. The same holds for the log-likelihood values. As can be observed in Table 4.4, the quality of the fit increases with the number of parameters in the model.

Model	Parameters	log-likelihood	BIC	$(-2/n)$BIC
E	10	−3645	−3685	2.3503
D	6	−3671	−3695	2.3566
C	5	−3677	−3697	2.3579
B	2	−4373	−4381	2.7940
A	1	−4659	−4664	2.9745

Table 4.4: Values of the log-likelihood and Bayesian information criterion for the fitted models without maintenance.

For the overall bridge conditions, model E fits best to the data and model A the least. Because the complexity of the model, and therefore also the computational effort, increases as more parameters are introduced, a test of the relative quality of fit can be performed. The purpose of this test is to determine if increasing the flexibility of the model, results in a statistically significant improvement of the fit. If this is not the case, then the decision maker may decide not to use the model with more parameters. It is important to note that all models have been fitted to the same data set, which means that the log-likelihood values in Table 4.4 may be compared to eachother. A slightly different data set will naturally result in a different log-likelihood value. For example, models which allow for backward transitions (e.g., due to maintenance) will be discussed later in this chapter and as these models allow for more data to be used, their log-likelihood values will be smaller compared to those in Table 4.4.

In order to test the statistical significance of the improvement in the quality of fit, the generalized ratio test may be applied. Let $L_A^* = L(\boldsymbol{\theta}_A^*; \boldsymbol{x})$ and $L_B^* = L(\boldsymbol{\theta}_B^*; \boldsymbol{x})$ be the likelihood functions for models A and B maximized by their respective optimal parameter values $\boldsymbol{\theta}_A^*$ and $\boldsymbol{\theta}_B^*$, then a well known theorem (Mood et al., 1974, Theorem 7 on p.440) states that

$$\kappa = -2\log\{L_A^*/L_B^*\} = -2\{\ell_A^* - \ell_B^*\}$$

has approximately a Chi-square probability distribution with degrees of freedom equal to the difference in the number of parameters between the two models. Loosely speaking, the hypothesis is that both models are

equally likely, or, that $\kappa = 0$. Given a significance level $0 < \alpha < 1$, we reject this hypothesis if $\kappa > \chi^2_{1-\alpha}(r)$, where $\chi^2_{1-\alpha}(r)$ is the $1-\alpha$ quantile of the Chi-square distribution with $r > 0$ degrees of freedom. This means that if the likelihood ratio of the estimated set of parameter values is in the top $100 \times \alpha$ percent out of all likelihood ratio values which can be obtained, the difference between the two models A and B is statistically significant. Using this test, it can be concluded that each model, starting with model A and ending with model E, represents a statistically significant improvement over the previous model. Even the small difference between models D and C is significant. In fact, model D is $e^8 \approx 3000$ times more likely to generate the observed data compared to model C.

An important remark must be made here: the fact that age-dependent Markov models fit better to the data compared to age-constant models, is not (necessarily) due to the underlying process being age-dependent. Incorporating age-dependent transition rates merely increases the number of parameters and therefore the flexibility of the model to adapt to the data. Therefore, this result does not constitute a proof of transitions in bridge condition data being age-dependent. In fact, the time-dependence in model D is not very strong and model E can only be implemented using a numerical approximation, like the Euler scheme described in Section 7.2.3, to calculate the transition probability function. From a practical point of view, model C is therefore a reasonable model to use for the deterioration process.

Also included in Table 4.4 are the values of the Bayesian information criterion (BIC) for each of the five models. This 'measure' for the quality of fit for a model M was initially proposed by Schwarz (1978) as $\ell_M(\boldsymbol{\theta}; \boldsymbol{x}) - (1/2)d_M \log(n)$, where $d_M > 0$ is the dimension of model M and n is the size of the data set $\boldsymbol{x}$. It avoids overfitting a model to data by taking into account the complexity of the model, which is reflected by its dimension. In essence, it adds a penalty for the number of parameters in the model. The Akaike information criterion (AIC) aims to do the same, but the BIC is to be preffered as it will 'consistently select the true model out of two contenders' (Lindsey, 1996). The last column of Table 4.4 contains the BIC values multiplied by $-2/n$, which is a formulation commonly found in the literature. Both formulations support the previous conclusions about the relative fit of the five models, but the last column in the table highlights the fact that the improvement is minimal for models beyond model C.

If the expected condition over time for model C is compared with the observed average conditions in the database, like in Figure 4.5, then the first conclusion may be that the model does not fit to the data very well at all. In fact, the model predicts the deterioration to be much faster than the data suggests. However, the largest number of observations are made at ages less than 10 years which mostly involves states 0 to 2. Transitions occur quite

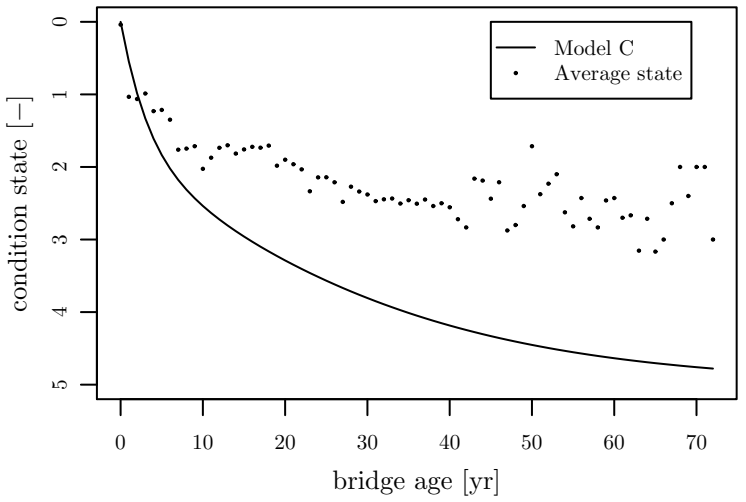


Figure 4.5: Expected condition state of model C compared to the average condition state as a function of bridge ages.

fast in these states, because a new structure will become damaged quite fast. The maximum likelihood estimator tries to follow these observations more closely compared to those at later ages due to the much larger number of observations. In the next section, a regression-based method will be used to estimate the parameters in model C and the result will be compared with the result presented here.

Comparison with results using regression onto observed states

It is interesting to compare the results obtained by maximum likelihood estimation, with those obtained by one of the more common estimation methods found in the literature, namely those using regression. Two formulations of a least squares regression were discussed in the section entitled "Regression using the state expectation" starting on page 26: a non-weighted least squares regression of the mean of the process $X(t)$ onto the observed states in Equation (2.11) and a weighted least squares regression of the state distribution $\boldsymbol{p}(t)$ onto the observed proportion of states $\boldsymbol{y}(t)$ in Equation (2.13). Using model C from Table 4.1, the regular least squares formulation of Equation (2.11) did not result in reasonable values for the parameters. The same holds for the weighted least squares formulation of Equation (2.13) if all weigths are removed (by setting $n(t) = 1$ for all t). In both cases, the transition intensity out of the initial state '0' is extremely high and the intensities out of the subsequent states are extremely low.

The regression using the weighted least squares formulation in Equation (2.13) used by Cesare et al. (1994), does converge to more realistic

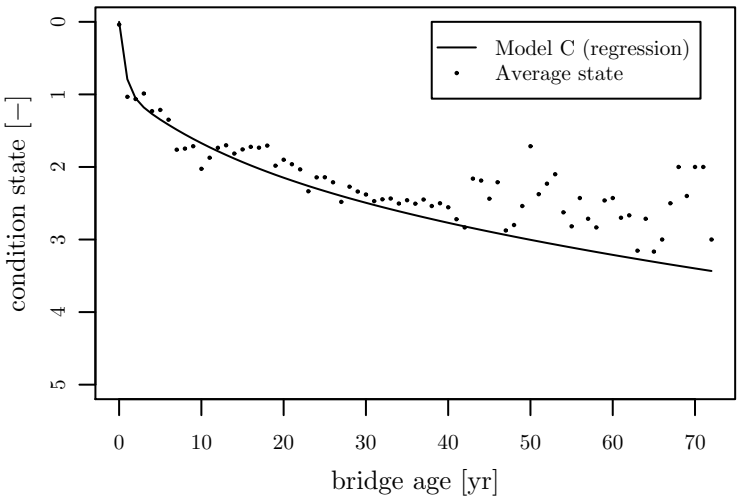


Figure 4.6: Expected condition state for model C fitted by regression onto the observed state distribution.

parameter values. These are $\boldsymbol{\lambda} = \{0.115, 0.008, 0.003, 0.001, 0.003\}$. The expectation of the model with these parameter values, is shown in Figure 4.6 and is visually more appealing compared to the result in Figure 4.5. The mean of the deterioration process follows the average observed condition quite closely, especially up to bridge ages of about 40 years for which the greatest number of observations are available. From this perspective, this approach may seem more appropriate for fitting Markov processes to bridge condition data. However, the log-likelihood of the data being generated by model C with the aforementioned parameter values is -5050, which is much smaller than the log-likelihood of -3677 obtained using the parameters in Table 4.2. Therefore the maximum likelihood approach results in a better fit from the perspective of transitions.

Also, the mean time to reach the absorbing state 5 from the initial state 0 is 142 years for the model with the parameters estimated by regression. The corresponding probability distribution is shown in Figure 4.7. The very long time to reach the final state is not realistic.

The primary objection raised against the regression models in Section 2.3, is that they do not account for the Markov property in the deterioration model. The result of this is clearly observable in Figure 4.6, which shows that the model tries to closely follow all observed states. However, the condition states observed for structures at an age above 35 to 40 years, are known to be less reliable because major renovation is often performed around these ages. Even if there is no record of maintenance available, it is very likely that maintenance is included in the observed condition

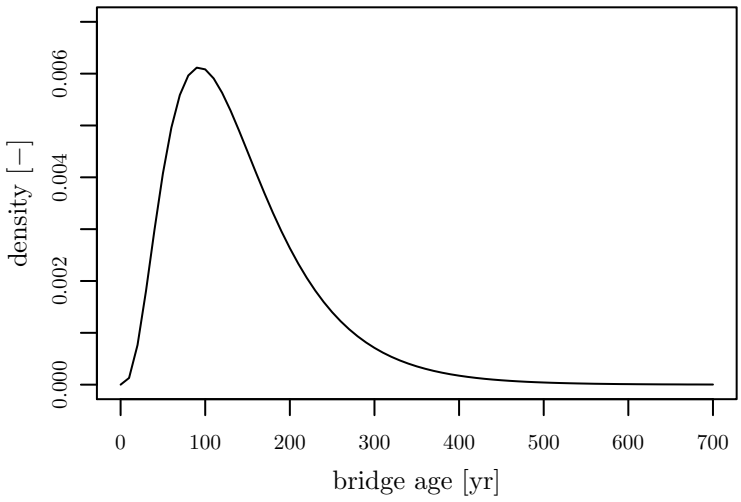


Figure 4.7: Probability of the time to reach the final state for model C which is fitted using regression onto the state distribution.

states. The most reliable information is therefore given by observations of structures of age less than 35 to 40 years. Since most condition data in the Netherlands belong to this category, the maximum likelihood estimation is primarily driven by these more reliable observations.

Distribution of the estimated parameters

The third problem of statistical modeling, as defined by Fisher (1922), is the problem of determining the probability distribution of the estimator. As mentioned in Chapter 3, the maximum likelihood estimator has the pleasant property that the estimator is asymptotically normal (or Gaussian). This means that the estimator tends to have a normal distribution if a sufficient number of samples were used in the estimation.

The robustness of the estimation procedure may be tested by performing a bootstrap. A bootstrap consists of sampling a large number of new data sets from the original data set and estimating the model parameters using these new data sets. Therefore, the bootstrap technique is just another way of determining the distribution of the estimator. If the procedure is robust, the estimated parameters for the model using these new data sets should be close to the estimated parameters of the original data set and they should be approximately normally distributed as explained above. Here, a bootstrap has been performed on the overall bridge condition data set by randomly (i.e., uniformly) sampling, with replacement, approximately 2300 condition sequences and thus creating 100 new data sets of the same size as the underlying data set. For model A, Figure 4.8 shows that the

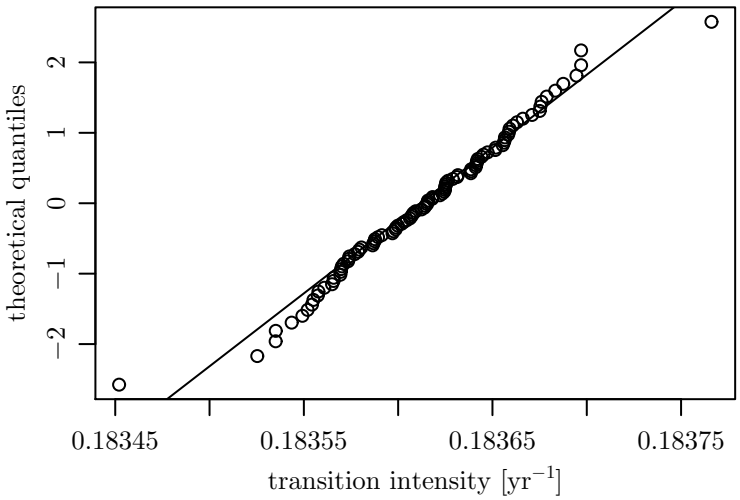


Figure 4.8: Normal probability plot of the bootstrapped transition intensity of model A.

estimated transition intensity of the bootstrapped data sets are very close to the transition intensity of the original data set, which is $\lambda \equiv a = 0.18$.

A simple test of normality indicates that the hypothesis of normality can not be rejected. The linearity of the data in Figure 4.8 shows that the sample distribution is close to being Gaussian, but has slightly longer tails.

The same bootstrapping technique can be applied to the estimation of the five parameters in model C. The result for all five parameters is shown in Figure 4.9, which shows that the estimates are close to the original estimate, but the distribution for all five parameters is not Gaussian. The distributions are more peaked compared to the Gaussian distribution, which can be observed in the histograms presented in Figure 4.10. This result may be due to the fact that the number of available samples is too small or that the accuracy of the estimation procedure is not high enough. As the estimates based on the bootstrapped data sets are very close to the estimates based on the original data set, even small rounding errors or insufficient computational accuracy may distort the resulting distribution. In any case, this result is quite good as it shows that the model and the estimation of the parameters are robust: there are no extreme differences in the resulting parameter estimates.

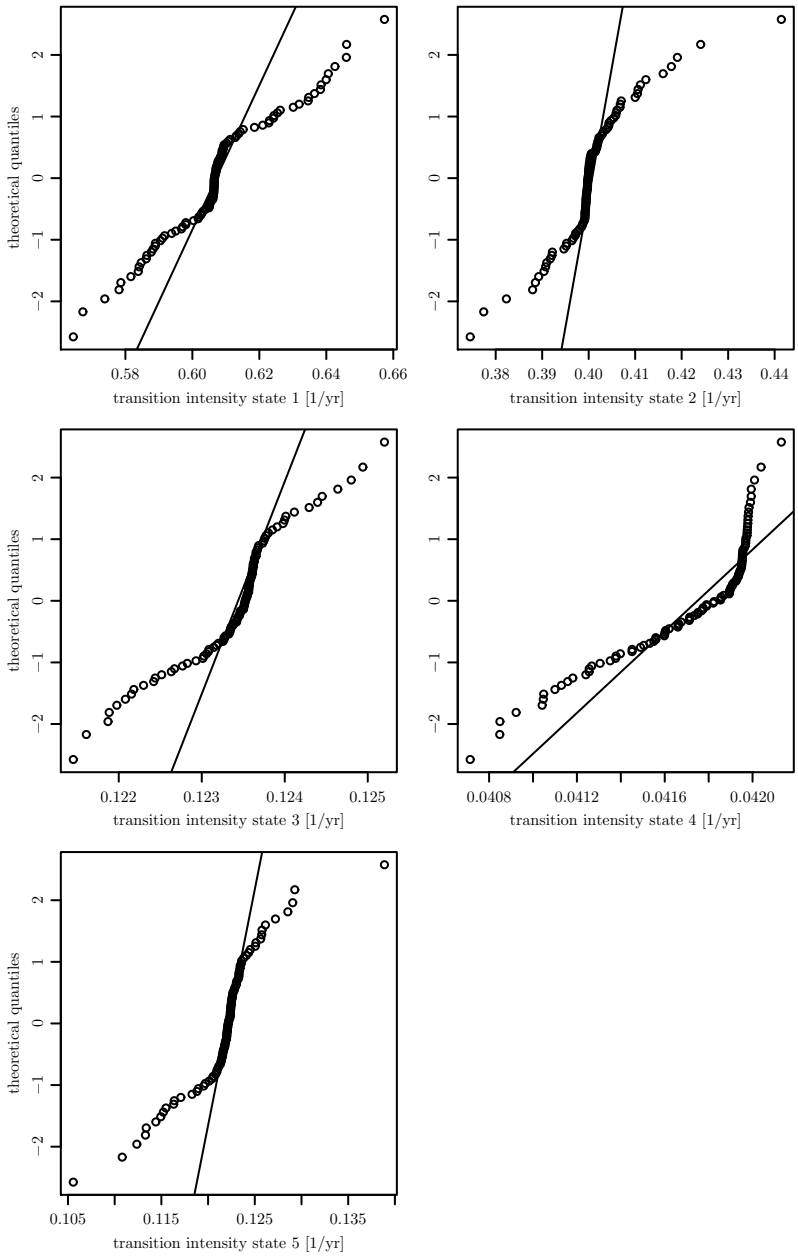


Figure 4.9: Normal probability plots of the five bootstrapped transition intensities in model C.

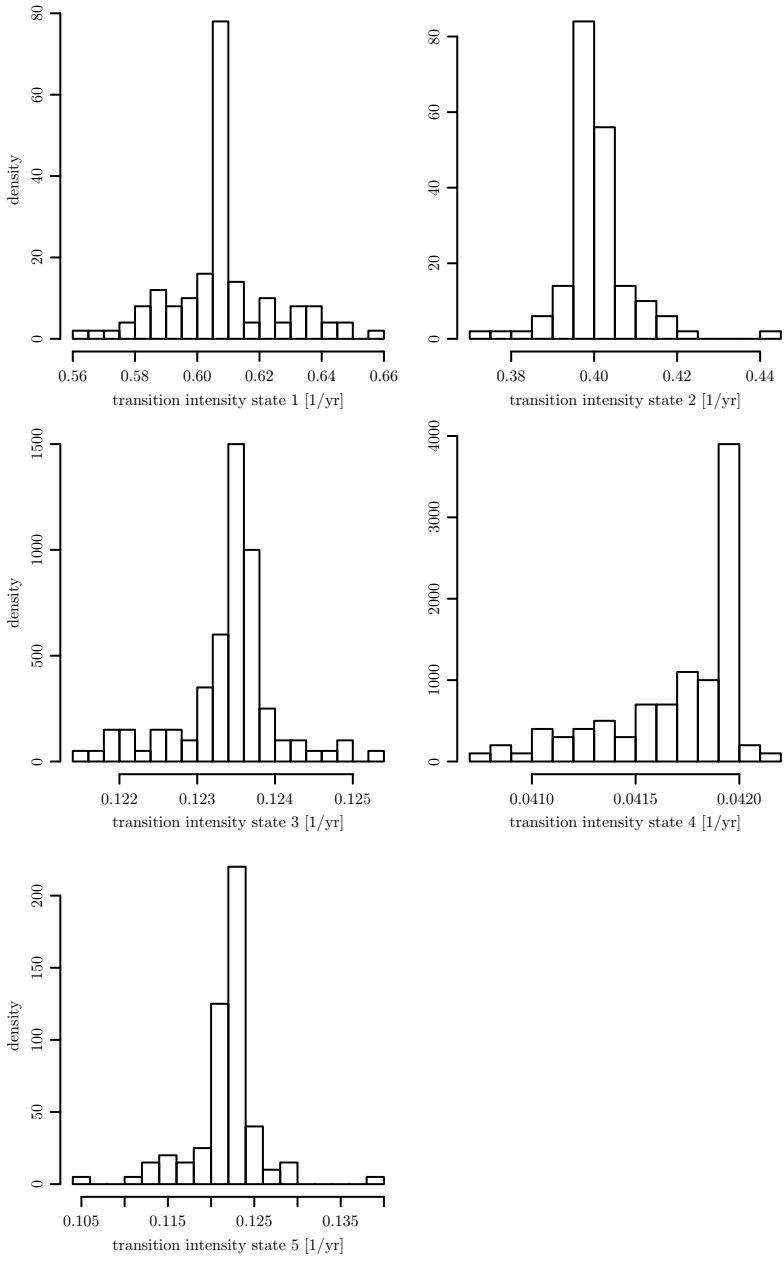


Figure 4.10: Histogram plots of the five bootstrapped transition intensities in model C.

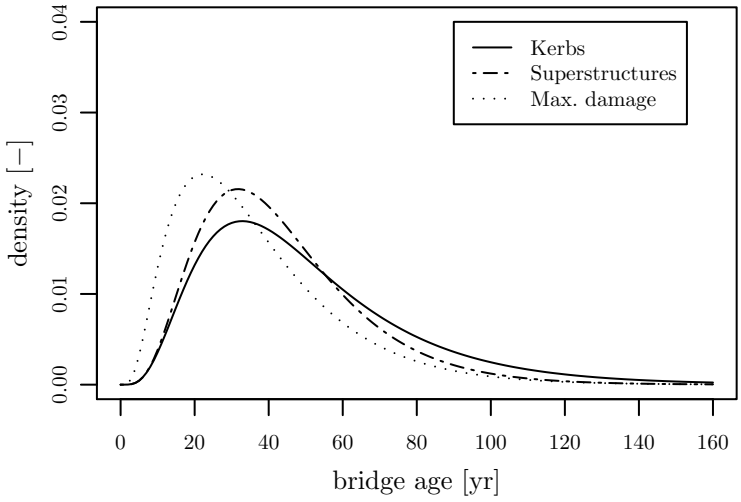


Figure 4.11: Probability density of the time to reach the final state for the maximum damage, superstructure and kerb condition data sets.

Maximum damage, superstructure and kerb conditions

Model C, with state-dependent and age-independent transition intensities, is also fitted to the available condition data for the maximum damage, superstructures and kerbs. The probability distribution of the time to reach the last state and the expected condition state as a function of bridge age, are presented in Figures 4.11 and 4.12 respectively.

As can be expected, structures deteriorate faster if the most severe damage is used as a representation of the condition. The mean time to reach state 5 is 37 years compared to 45 (see Table 4.3) when using the overall bridge conditions. The results in Table 4.5 show that the transition intensity out of the initial state 0 is particularly high at more than two per year. This means that it generally takes less than half a year for at least one damage of severity 1 to appear. The condition development of superstructures and kerbs is quite similar, although the kerbs have significantly more uncertainty in the higher values for the time to reach state 5. This also results in a higher expectation for the time to reach state 5, namely 49 years compared to 43 years for superstructures.

The 90% confidence bounds in the results for superstructures and kerbs are not noticably narrower compared to those of the overall bridge conditions shown in Table 4.3. Although it might have been expected that these bounds would be narrower due to the fact that there is less variability in the source of the conditions (i.e., the overall bridge condition data is based on

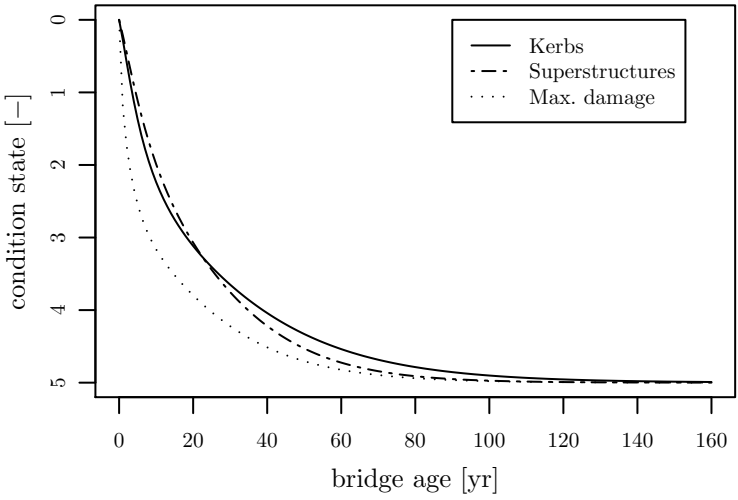


Figure 4.12: Expected condition state as a function of age for the maximum damage, superstructure and kerb condition data sets.

Data	a_0	a_1	a_2	a_3	a_4	5% [yr]	Mean [yr]	95% [yr]
Max. damage	2.168	0.605	0.290	0.058	0.076	10	37	79
Superstruc.	0.198	0.394	0.118	0.062	0.092	16	43	85
Kerbs	0.248	0.564	0.121	0.040	0.095	16	49	105

Table 4.5: Estimated parameters, the mean value and the 5% and 95% percentiles of the time to reach the final state for the maximum damage, superstructure and kerb data sets.

all components in the structure, whereas superstructures consist of fewer components and kerbs are themselves assumed to be a component), this can not be supported by these results.

4.2.2 Models with backward transitions or maintenance

In this section two models with backward transitions will be fitted to the overall bridge condition. The first model allows for transitions back to the previous state, refered to as model F, and the second model allows for transitions back to state 0, which is referred to as model G. The latter model represents a transition structure in which perfect maintenance is possible. Because deterioration should always be separated from maintenance (such that they can be treated separately in the decision model), the purpose

of this section is simply to observe the behaviour of the data with respect to backward transitions. Because there are no detailed records of maintenance in the database, it may be possible to observe transitions due to maintenance in the data.

The transition intensity matrix for model F which has the lowest log-likelihood value is:

$$\boldsymbol{Q}_a = \begin{bmatrix} -1.572 & 1.572 & 0 & 0 & 0 & 0 \\ 0.216 & -0.612 & 0.396 & 0 & 0 & 0 \\ 0 & 0.120 & -0.408 & 0.300 & 0 & 0 \\ 0 & 0 & 0.336 & -0.552 & 0.216 & 0 \\ 0 & 0 & 0 & 6.228 & -6.348 & 0.120 \\ 0 & 0 & 0 & 0 & 0.120 & -0.120 \end{bmatrix}.$$

Here, the high intensity of backward transitions from state 4 to state 3 is most obvious. In general, structures in states 3 and 4 are more likely to transition backwards compared to structures in states 1 and 2. This possibly reflects an increase in the difference of opinion between inspectors around states 3, 4 and 5. This result has a log-likelihood value of -4777, which is necessarily lower than the results in Table 4.4, because more data is used by this model.

Another result for model F is the transition intensity matrix given by:

$$\boldsymbol{Q}_b = \begin{bmatrix} -0.372 & 0.372 & 0 & 0 & 0 & 0 \\ 0.072 & -0.312 & 0.240 & 0 & 0 & 0 \\ 0 & 0.072 & -0.228 & 0.156 & 0 & 0 \\ 0 & 0 & 0.156 & -0.144 & 0.096 & 0 \\ 0 & 0 & 0 & 0.384 & -0.468 & 0.096 \\ 0 & 0 & 0 & 0 & 0.192 & -0.192 \end{bmatrix}.$$

The model with this transition intensity matrix has a lower log-likelihood value of -5224, but visually it fits closer to the average of the observed conditions. This can be observed in Figure 4.13, where model F(a) uses $\boldsymbol{Q}_a$ and model F(b) uses $\boldsymbol{Q}_b$. Although model F(b) is less likely to generate the data, the parameters in $\boldsymbol{Q}_b$ are less extreme compared to those in $\boldsymbol{Q}_a$. The effect of more transitions going backwards, if starting from states 3 and 4, is also present in $\boldsymbol{Q}_b$.

When comparing the indicator of the most severe damage with the overall bridge condition, it can be observed that inspectors are more likely to change the default condition assignment. If there is one damage which in itself is quite severe, but not representative for the overall condition of the structure, the inspector will not assign the severity of the individual damage to the whole structure. In theory, this means that inspectors should do this for all damage severities, but as was discussed in Kallen and van Noortwijk (2006b), this is not the case. If a damage of severity 2 is found,

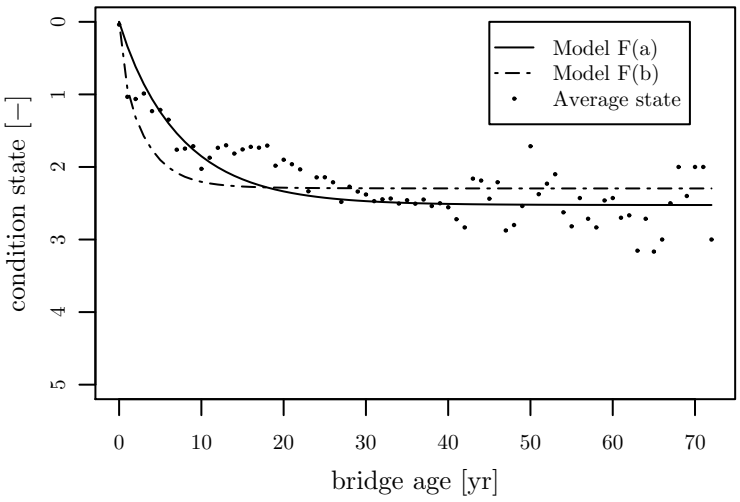


Figure 4.13: Expected condition state for model F, compared to the average state in the data, using two sets of estimated parameter values.

but the damage is not representative for the structure, the inspectors are less inclined to change the default condition assignment compared to if the damage had a severity 3 or higher. This phenomenon is likely to be at least one source of the backward transitions observed in model F.

Next, a model with perfect maintenance is considered. In this case, objects are allowed to transition back to the initial state. The estimated transition intensity matrix for this model, referred to as model 'G', is

$$\boldsymbol{Q} = \begin{bmatrix} -13.139 & 13.139 & 0 & 0 & 0 & 0 \\ 3.155 & -4.041 & 0.886 & 0 & 0 & 0 \\ 0.069 & 0 & -0.283 & 0.214 & 0 & 0 \\ 0.146 & 0 & 0 & -0.266 & 0.120 & 0 \\ 1.326 & 0 & 0 & 0 & -1.481 & 0.155 \\ 0.426 & 0 & 0 & 0 & 0 & -0.426 \end{bmatrix}.$$

The expectation of the condition state as a function of the age of a structure is shown in Figure 4.14. The extremely high intensities out of state 0 and from state 1 back to state 0 are not very realistic. Nonetheless, this model has a log-likelihood of -4799, which is close to the likelihood of model F with transition intensity matrix $\boldsymbol{Q}_a$.

It is not unusual for the Dutch bridge condition data to result in fast transitions out of the initial states. This is because the overall bridge condition is very much like a series system in which the condition of the structure

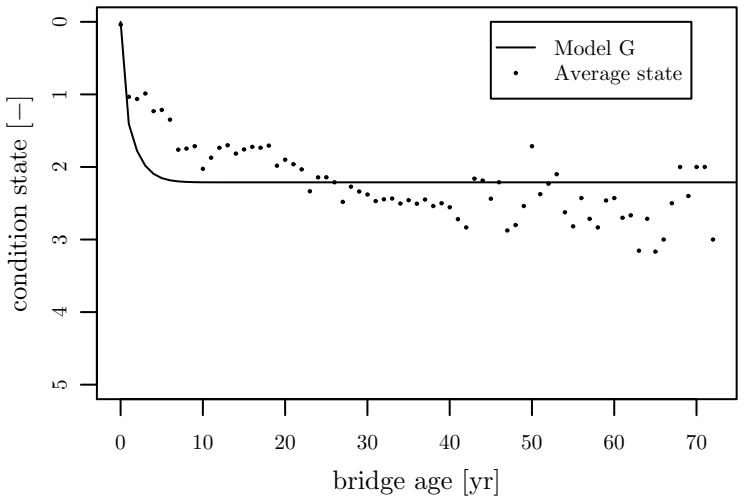


Figure 4.14: Expected condition state for model G, compared to the averag state in the data.

is the same as the condition of the weakest link or element. As soon as a single damage is found, the condition of the structure will transition out of the initial 'perfect' state. This is of course due to the automatic assignment of the most severe damage indicator to the condition of the structure; see Section 1.3.

4.3 Inclusion of inspection variability

The probabilities of observing a state $Y_k = j$ given that the true state is i at the k-th observation for each possible pair of states i and j, can be collected in an error or misclassification matrix:

$$E = \begin{bmatrix} e_{00} & e_{01} & e_{02} & e_{03} & e_{04} & e_{05} \\ e_{10} & e_{11} & e_{12} & e_{13} & e_{14} & e_{15} \\ e_{20} & e_{21} & e_{22} & e_{23} & e_{24} & e_{25} \\ e_{30} & e_{31} & e_{32} & e_{33} & e_{34} & e_{35} \\ e_{40} & e_{41} & e_{42} & e_{43} & e_{44} & e_{45} \\ e_{50} & e_{51} & e_{52} & e_{53} & e_{54} & e_{55} \end{bmatrix}, \tag{4.2}$$

where $e_{ij} = \Pr\{Y_k = j \,|\, X_k = i\}$, therefore $0 \leq e_{ij} \leq 1$ for all i and j, and $\sum_j e_{ij} = 1$. Note that e_{ii} is actually the probability of correctly identifying state i as the current condition. Only e_{ij} for $i \neq j$ represents a true misclassification. It is assumed here that the probabilities of misclassification do not change over time. Also, the error matrix is the same for all inspections,

therefore no distinction is made between different inspectors, weather condition at the time of the inspection, or any other factor which may influence the judgment of the inspector. As with the transition intensity matrix, it is possible to 'design' the model by selecting a particular structure for the misclassification matrix. For example, if the inspectors are assumed to be wrong by only one state, then the misclassification matrix is chosen to have the form

$$E = \begin{bmatrix} e_{00} & e_{01} & 0 & 0 & 0 & 0 \\ e_{10} & e_{11} & e_{12} & 0 & 0 & 0 \\ 0 & e_{21} & e_{22} & e_{23} & 0 & 0 \\ 0 & 0 & e_{32} & e_{33} & e_{34} & 0 \\ 0 & 0 & 0 & e_{43} & e_{44} & e_{45} \\ 0 & 0 & 0 & 0 & e_{54} & e_{55} \end{bmatrix}. \tag{4.3}$$

Once the structure of the misclassification matrix has been selected, there are two options for quantifying the error probabilities: either they are assessed by expert judgment or their most likely values are estimated together with the transition intensities of the underlying Markov deterioration process. Both options are discussed here, starting with the latter. If there are n states, the error matrix (4.3) requires $2(n-1)$ parameters to be estimated and the error matrix (4.2) adds $n(n-1)$ parameters to the maximum likelihood estimation procedure. For the Dutch bridge conditions there are six states, which means that respectively ten and thirty probabilities would have to be estimated. As this significantly increases the number of parameters to be estimated in the model and therefore also the computational effort, this is not an attractive approach. An alternative approach is to select a finite and discrete probability distribution for each row of the error matrix E. As an example, the binomial probability distribution, given by

$$e_{ij} = \binom{n-1}{j} \phi^j (1-\phi)^{n-j-1}, \tag{4.4}$$

where $j = 0, 1, \ldots, n-1$ and $0 \leq \phi \leq 1$. The expectation for the to be observed state j, given the true state i, is $(n-1)\phi$, where ϕ is the parameter for the probability distribution in each row i. This approach requires only one parameter for each row (i.e., six parameters for the Dutch condition data), but allows for less flexibility because the shape of the binomial can not take on any arbitrary form. The result for the misclassification matrix Equation (4.2) using the binomial distribution for each row, is shown in Figure 4.15.

The expectation of the observation process $Y(t)$, $t \geq 0$ is calculated in a similar way as the expectation of the true states $X(t)$:

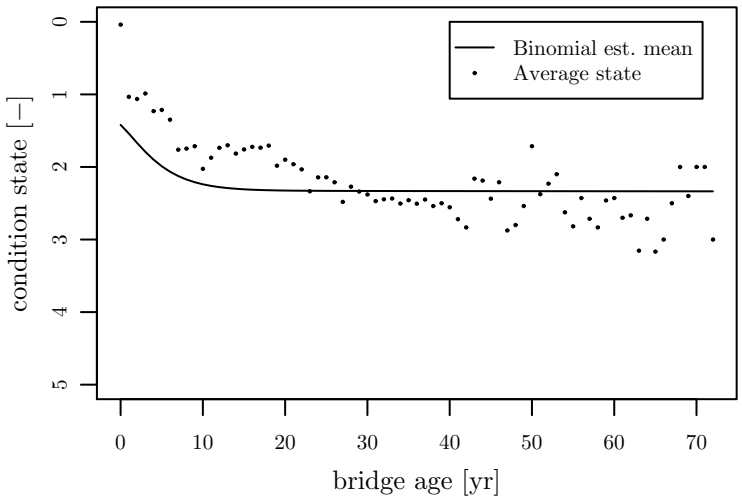


Figure 4.15: Expected condition state as a function of age for the hidden Markov model fitted to the overall bridge condition data.

$$\mathbb{E}Y(t) = \sum_j j \sum_i \Pr\{Y(t) = j \mid X(t) = i\} \Pr\{X(t) = i\}.$$

The result in Figure 4.15 looks similar to those in Figure 4.13. Due to the variability in the inspections, the expected condition state may start in a state other than state 0. The estimated error matrix is

$$E = \begin{bmatrix} 0.19 & 0.37 & 0.30 & 0.12 & 0.02 & 0.00 \\ 0.13 & 0.32 & 0.33 & 0.17 & 0.04 & 0.01 \\ 0.04 & 0.19 & 0.33 & 0.29 & 0.13 & 0.02 \\ 0.03 & 0.15 & 0.31 & 0.32 & 0.16 & 0.03 \\ 0.15 & 0.35 & 0.32 & 0.14 & 0.03 & 0.01 \\ 0.04 & 0.19 & 0.33 & 0.29 & 0.13 & 0.02 \end{bmatrix}$$

The second option for quantifying the probabilities of misclassification, is to use expert judgment. In Sztul (2006) an informal expert judgment approach was applied. For the full error matrix Equation (4.2), two probability distributions were chosen with a fixed mean: the binomial and the maximum entropy distributions. The basic assumption was that the expert is expected to correctly identify the true state, therefore the mean of each row distribution was set to the true state of that row. For the binomial distribution, this results in an error matrix

$$E = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.33 & 0.41 & 0.20 & 0.05 & 0.01 & 0.00 \\ 0.08 & 0.26 & 0.34 & 0.23 & 0.08 & 0.01 \\ 0.01 & 0.08 & 0.23 & 0.34 & 0.26 & 0.08 \\ 0.00 & 0.01 & 0.05 & 0.20 & 0.41 & 0.33 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

and for the maximum entropy distribution this results in

$$E = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.48 & 0.25 & 0.13 & 0.07 & 0.04 & 0.03 \\ 0.25 & 0.21 & 0.17 & 0.15 & 0.12 & 0.10 \\ 0.10 & 0.12 & 0.15 & 0.17 & 0.21 & 0.25 \\ 0.03 & 0.04 & 0.07 & 0.13 & 0.25 & 0.48 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}.$$

The maximum entropy distribution is a discrete probability distribution which adds the least information (or which has maximum entropy) with a given mean; see Jaynes (1957). It is less 'informative' than the binomial distribution with given mean. The purpose of using the maximum entropy distribution is to add as little extra information as possible in order to obtain a result which is not too 'colored' by the choice of the error distribution. Note that for both approaches, the first and last state are perfectly observable by the inspector. Another noticable effect of this approach is that the inspectors would tend to underestimate the condition in the initial true states (i.e., states 1 and 2), whereas they would tend to overestimate the condition in later states (i.e., states 3 and 4). From practical experience in the Netherlands, this should be exactly the opposite in practice. See Sztul (2006) for the results of these two approaches.

4.4 Analysis of covariate influence

In Section 3.2.3, a short introduction to covariate analysis was given. The purpose of covariate analysis is to determine if a certain grouping of structures, according to various characteristics, would result in significantly different transition intensities. If this is the case, then the decision maker may choose to treat these groups separately in the maintenance model. In this section, a covariate analysis will be performed on the overall condition data from bridges in the Netherlands.

First, model A is considered with three covariates: whether a bridge is located in or over the motorway, whether it is designated as a bridge or viaduct, and whether the structure is located in a province with a high or low population intensity. In the Netherlands, a 'bridge' is generally defined as a structure over water and a 'viaduct' as a structure which connects

any two points in order to cross over an obstruction like, for example, a motorway. The log-linear model is then given by

$$\lambda = \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\},$$

where

$$X_1 = \begin{cases} 0, & \text{if located in the motorway, and} \\ 1, & \text{if located over the motorway.} \end{cases}$$

$$X_2 = \begin{cases} 0, & \text{if designated as a concrete viaduct, and} \\ 1, & \text{if designated as a concrete bridge.} \end{cases}$$

$$X_3 = \begin{cases} 0, & \text{if located in a more densily populated province, and} \\ 1, & \text{if located in a lesser populated province.} \end{cases}$$

The null-hypothesis for each covariate is that it does not significantly influence the model outcome; that is $H_0 : \beta_i = 0$. This hypothesis should be rejected if the coefficient β_i is significantly different from zero. The results for model A are collected in Table 4.6. The last column in the table shows if the null-hypothesis is rejected or not. It is only rejected for β_3 (even at the 1% significance level), therefore this covariate (or independent variable) is significant. The other two covariates may influence the result as well, but from a statistical point of view this can not be shown with certainty. The null-hypothesis may also be rejected for coefficient β_1 if considering only a 10% significance level, but this level is rarely used.

i	$\hat{\beta}_i$	$\mathrm{SE}_{\hat{\beta}_i}$	p-value	reject
0	−4.2470	0.0291	n/a	n/a
1	0.0584	0.0388	0.07	no
2	−0.0144	0.0475	0.38	no
3	0.1529	0.0366	0.00	yes

Table 4.6: Estimated coefficients and their standard error and corresponding p-value for the covariates in model A.

The base transition intensity is determined by the intercept β_0. The base transition intensity per month is $\exp\{-4.2470\} = 0.0143$ or approximately 0.17 per year. If the structure is located in a more densily populated area (i.e., $X_3 = 1$) and does not possess the two other properties (i.e., $X_1 = X_2 = 0$), then this base intensity is multiplied with a factor $\exp\{0.1529\} = 1.1652$ per month and the annual transition intensity would be 0.20.

As mentioned in Chapter 3, the maximum likelihood estimator is asymptotically normal and in Section 3.2.3 it was mentioned that Wald's test essentially rejects the hypothesis that $\hat{\beta}_i = 0$ if zero is not within the $(1-\alpha)\%$ bounds. In Figure 4.16 it can be clearly observed how this results in the hypothesis $\hat{\beta}_3 = 0$ being rejected, whereas the others are not.

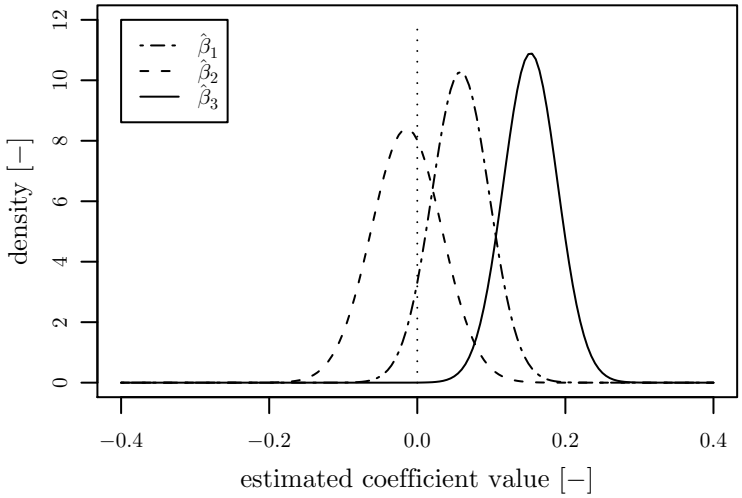


Figure 4.16: Asymptotic density of the coefficients for the explanatory variables in model A.

SUMMARY

The primary research goal was to see if Markov processes could be used to model deterioration of structures in the Netherlands using the condition data which is available and to determine how this could best be done. This chapter describes the results of the maximum likelihood estimation procedure introduced in the previous chapter.

The properties of the various data sets and how these were obtained from the database, is the topic of Section 4.1. Three data sets have been analysed: overall bridge condition, condition of superstructures and of kerbs. These data sets were extracted from the relational database used in the Netherlands for the registration of inspection results. The database was not designed with a statistical analysis in mind, therefore the extraction of the data and conversion into a suitable format for analysis requires special care and a great deal of time.

For the models which allow only for deterioration (i.e., transitions to worse states), the quality of the fit onto the data increases as the number of parameters increases. Age-dependent or inhomogeneous processes therefore tend to fit better due to the fact that they are more flexible to adapt to the data and not necessarily because the data is age-dependent by nature. From a practical point of view, model C, which has state-dependent transition intensities (i.e., they may be different for each state), is the most appealing model, even if others have a higher likelihood of generating the given data. Optionally, model D may also be used. It allows for an overall change in

the transition intensities as a function of the structure's age and requires only slightly more calculations to be made. However, the results indicate that this change over time is not very strong, i.e. the intensities are close to constant throughout a structure's life.

If the possibility of backward transitions is included in the deterioration process, to account for inspection variability or maintenance activities, the general trend is that structures are more likely to move towards a worse state in the initial states 0, 1 and 2, but the reverse is observed for states 3, 4 and 5.

The inclusion of inspection variability, by use of hidden Markov models, has been deemed not to be useful for application in the Netherlands. The probabilities of misclassification must be estimated from the data or determined by expert judgment. Either way, by definition of visual inspections, the actual process can never be observed and therefore this process is completely determined by the subjective choice for the misclassification probabilities. For the Dutch bridge condition data, both the subjective estimation approach, and the data-based estimation, did not lead to satisfactory results. From a decision maker's point of view, the subjective estimation approach is most likely to be preferred, as the data-based estimation may not result in a realistic misclassification structure. Also, the data-based estimation introduces many more parameters to be estimated, which may pose a problem if the amount of data is insufficient.

Finally, a covariate analysis was performed to determine the significance of the influence of independent variables. These independent variables describe if a structure possesses a certain characteristic or not. In this analysis, three independent variables were considered: whether a bridge is located within the motorway or over it, whether the structure is designated to be a bridge or a viaduct and if the structure is located in a province with a relatively high population density. Only the latter has a statistically significant influence, therefore separate maintenance policies may be used for structures in different locations throughout the country.

5

Optimal maintenance decisions

Once the deterioration model, using a finite-state Markov process, has been chosen and fitted to the available data, it can be used to make a decision about how often a structure should be inspected and which type of maintenance should be performed at what time. This chapter therefore deals with the decision model as discussed in Chapter 1 and shown in Figure 1.1.

For finite-state Markov processes, there are two decision models which may be used: a Markov decision process or a condition-based maintenance model. Markov decision processes are most commonly applied as they are popular in general (Dekker, 1996). A short introduction to the theory of Markov decision processes is given in Section 5.1. The condition-based maintenance model is presented in Section 5.2 which is largely based on Kallen and van Noortwijk (2006a). A comparison between both models is given, together with a discussion on the application of the condition-based maintenance model in the Netherlands, in the summary at the end of this chapter. The model presented here only considers the economics of inspections and maintenance. In practice, thisis just one, albeit a very important one, of many aspects to consider when deciding on a maintenance policy.

5.1 MARKOV DECISION PROCESSES

In order to be able to make decisions about an optimal policy for maintenance actions, a finite set of actions A and costs $C(i,a)$ have to be introduced, which are incurred when the process is in state i and action $a \in A$ is taken. The costs are assumed to be bounded and a policy is defined to be any rule for choosing actions. When the process is currently in state i and an action a is taken, the process moves into state j with probability

$$p_{ij}(a) = \Pr\{X(n+1) = j \mid X(n) = i, a_n = a\}.$$

This transition probability again does not depend on the state history. If a stationary policy is selected, then this process is called a Markov Decision Process (MDP). A stationary policy arises when the decision for an action only depends on the current state of the process and not on the time at which the action is performed.

Now that the state of the structure over time with or without performing maintenance actions can be modeled, the optimization of inspection and maintenance policies using this process can be performed. For example, when the system is in state i, the expected discounted costs over an unbounded horizon are given by the recurrent relation

$$V_\alpha(i) = C(i,a) + \alpha \sum_{j=1}^{N} p_{ij}(a) V_\alpha(j), \tag{5.1}$$

where α is the discount factor for one year and V_α is the value function using α. This discount factor is defined as $\alpha = (1 + r/100)^{-1}$ with r the yearly discount percentage. Starting from state i, $V_\alpha(i)$ gives us the cost of performing an action a given by $C(i,a)$ and adds the expected discounted costs of moving into another state one year later with probability $p_{ij}(a)$. The discounted costs over an unbounded horizon associated with a start in state j are given by $V_\alpha(j)$, therefore Equation (5.1) is a recursive equation. The choice for the action a is determined by the maintenance policy and also includes no repair.

A cost optimal decision can now be found by minimizing Equation (5.1) with respect to the action a. A classic approach is to use a policy improvement algorithm, which consists of successively adjusting the policy until no further improvement can be made. Under some minor assumptions, this optimization problem can also be formulated as a linear programming problem, which can then be solved using the simplex algorithm. See Ross (1970, Chapter 6) for the theory of Markov decision processes.

A Markov decision process and the linear programming formulation was used in the Arizona pavement management system, see Golabi et al. (1982) and Golabi (1983), and in Pontis, see Golabi and Shepard (1997), amongst others. More references are given in the review by Frangopol et al. (2004).

5.2 CONDITION-BASED INSPECTION AND MAINTENANCE MODEL

The Markov decision processes discussed in the previous section, are essentially condition-based decision models as the decisions depend on the current state of the process. The model which will be discussed in this section is commonly referred to simply as the 'condition-based inspection and maintenance model' for stochastic processes, hence the title of this section.

5.2.1 MODEL

As discussed in Chapter 1, each maintenance model can be roughly separated into a deterioration model and a decision model. For the application

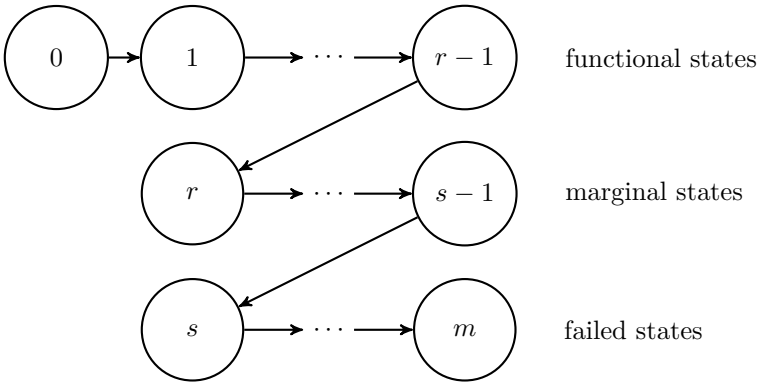


Figure 5.1: Graphical representation of the marginal checking model.

of the decision model, the structure of the Markov deterioration process is defined first.

Deterioration model

For the deterioration model, a sequential Markov process is used. A new object starts in state 0 and successively degrades through subsequent states as is shown in Figure 5.1. States r and s are the threshold states for preventive and corrective maintenance respectively. An object which has not yet reached state r is functional and if it has reached state r, but not yet state s, it is called marginal. States s and beyond are failed states. The term 'marginal' was used in the same context by Flehinger (1962) and indicates that the object is still functional, but is ready for a preventive repair or replacement.

Decision model

The decision model is based on renewal theory. It is therefore assumed that preventive and corrective repairs bring the object to an as-good-as-new state. The time between the service start of an object and its renewal represents a single life cycle. The renewal reward process $R(t)$ is defined as

$$R(t) = \sum_{n=1}^{N(t)} Y_n,$$

where Y_n is the reward earned at the n-th renewal. In the context of maintenance, the rewards are the costs of inspection and maintenance actions. The goal of the decision model is to minimize the long-term expected average costs per unit of time given by $\lim_{t\to\infty} \mathbb{E}R(t)/t$. According to the renewal reward theorem,

$$\lim_{t\to\infty} \frac{\mathbb{E}R(t)}{t} = \frac{\mathbb{E}C}{\mathbb{E}I}, \tag{5.2}$$

where C are the uncertain costs per cycle and I is the uncertain duration of a cycle; see for example Ross (1970, p.52). The expected average costs per unit of time over an unbounded horizon are therefore given by the ratio of the expected total costs incurred during the life cycle over the expected length of a life cycle. It is also possible to consider discounted costs in order to account for inflation over longer periods of time. However, for ease of presentation, discounting is not considered here.

Let $\tau > 0$ denote the fixed duration between inspections, such that $k\tau$ ($k = 1, 2, \ldots$) is the time of the k-th inspection. This is the primary decision variable. Other decision variables like the threshold state r for preventive maintenance may be included. Also, the time to the first inspection, as suggested by Jia and Christer (2002), is a useful decision variable. In this case, the inspections are performed at times $\tau_1 + (k-1)\tau$ for $k = 1, 2, \ldots$ and the optimal combination of τ_1 and τ, resulting in the lowest expected average costs per unit of time, are determined.

At each inspection, the object can be in a functional, marginal or failed state. If the object is found to be in a functional state, no maintenance is required and only the cost of the inspection is incurred. If it is found to be in a marginal state, the costs of preventive maintenance are added to the cost of the inspection. For an object in a failed state, there are two scenarios: either the failure is immediately noticed at the time of occurrence, without the necessity of an inspection, or failure is only detected at the next planned inspection. In the first case, only the costs of corrective maintenance are incurred and in the second case, the costs of the inspection and unavailability have to be included as well.

The costs per cycle are the sum of the costs of all inspections during the cycle and either a single preventive or a single corrective replacement. For the first scenario, in which failure is immediately noticed, the expected cycle costs are

$$\begin{aligned}\mathbb{E}C = \sum_{k=1}^{\infty} \Big[&(kc_I + c_R)\Pr\{\text{PR in } ((k-1)\tau, k\tau]\} + \cdots \\ &+ ((k-1)c_I + c_F)\Pr\{\text{CR in } ((k-1)\tau, k\tau]\}\Big],\end{aligned} \tag{5.3}$$

where PR is preventive repair, CR is corrective repair, and c_I, c_P and c_F are the costs of an inspection, preventive repair and a corrective repair respectively. The expected cycle length is

$$\mathbb{E}I = \sum_{k=1}^{\infty} \Big[k\tau \Pr\{\text{PR in } ((k-1)\tau, k\tau]\} + \cdots + \sum_{n=(k-1)\tau+1}^{k\tau} n \Pr\{\text{CR in } (n-1, n]\} \Big]. \tag{5.4}$$

The summation over n from $(k-1)\tau + 1$ to $k\tau$ reflects the immediate identification of a failure.

For the second scenario, in which failure is not noticed until the next inspection, Equations (5.3) and (5.4) become

$$\mathbb{E}C = \sum_{k=1}^{\infty} \Big[(kc_I + c_R) \Pr\{\text{PR in } ((k-1)\tau, k\tau]\} + \cdots + (kc_I + c_F) \Pr\{\text{CR in } ((k-1)\tau, k\tau]\} + \cdots + \sum_{n=(k-1)\tau+1}^{k\tau} c_U(k\tau - n) \Pr\{\text{failure in } (n-1, n]\} \Big], \tag{5.5}$$

and

$$\mathbb{E}I = \sum_{k=1}^{\infty} \Big[k\tau \Pr\{\text{PR in } ((k-1)\tau, k\tau]\} + \cdots + k\tau \Pr\{\text{CR in } ((k-1)\tau, k\tau]\} \Big], \tag{5.6}$$

where c_U are the costs of unavailability per unit time. This cost is added in Equation (5.5) as a penalty for leaving a structure in a failed state. The costs increase proportionally to the time that the object is left in the failed state. Without this cost, the cheapest solution would be to not inspect at all, because the average costs per unit of time will decrease as the cycle length increases.

5.2.2 Probabilities of preventive and corrective maintenance

Special case

In this section, the most simple type of Markov process, namely the state-independent and time-homogeneous Markov process is considered, which corresponds to model A as defined in Section 4.2.1. In this case, the waiting times are modeled by identical exponential distributions, that is, $T \sim \text{Exp}(\lambda)$ with $F_T(t) = 1 - \exp\{-\lambda t\}$. Let $S_n = \sum_{i=1}^{n} T_i$ be the time at which the n-th transition takes place, then the following equivalence holds:

$$X(t) \leq n \Leftrightarrow S_n \geq t$$

In words: if the state of the object is less than or equal to n at time t, the time required to perform n transitions is greater than or equal to t. Note that for a component with sequential condition states as depicted in Figure 5.1, S_n is also the first passage time of state n. Using this relationship, the probability of preventive repair becomes

$$\Pr\{\text{PR in } ((k-1)\tau, k\tau]\} = \Pr\{(k-1)\tau < S_r \leq k\tau, S_s > k\tau\}.$$

This is the probability that the threshold state r for preventive maintenance is first reached between the previous inspection at time $(k-1)\tau$ and the current inspection at time $k\tau$, and the threshold state s for corrective maintenance has not been reached before the current inspection. The analytical solution of this probability is

$$\begin{aligned}\Pr\{(k-1)\tau < S_r \leq k\tau, S_s > k\tau\} = \\ \sum_{j=0}^{s-r-1} \frac{(\lambda k\tau)^{j+r}}{(j+r)!} e^{-\lambda k\tau} \left[1 - I_{1-\frac{1}{k}}(r, j+1)\right],\end{aligned} \tag{5.7}$$

where

$$I_x(a,b) = \int_{\phi=0}^{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \phi^{a-1}(1-\phi)^{b-1} \mathrm{d}\phi$$

is the incomplete beta function for $0 \leq x \leq 1, a > 0$ and $b > 0$, see for example Abramowitz and Stegun (1965). The result in Equation (5.7) is obtained by splitting the probability and using the independence of the increments S_r and $S_s - S_r$:

$$\begin{aligned}&\Pr\{(k-1)\tau < S_r \leq k\tau, k\tau - S_r < S_s - S_r\} \\ &\quad = \Pr\{S_r \leq k\tau, k\tau - S_r < S_s - S_r\} - \cdots \\ &\qquad - \Pr\{S_r \leq (k-1)\tau, k\tau - S_r < S_s - S_r\}.\end{aligned}$$

Denoting the difference of these two probabilities by $A - B$, each of these can be calculated as follows:

$$
\begin{aligned}
A &= \int_{\phi=0}^{k\tau} \int_{\theta=k\tau-\phi}^{\infty} f_{S_r}(\phi) f_{S_s-S_r}(\theta) \mathrm{d}\theta d\phi \\
&= \int_{\phi=0}^{k\tau} f_{S_r}(\phi) \left[\int_{\theta=k\tau-\phi}^{\infty} f_{S_s-S_r}(\theta) \mathrm{d}\theta \right] \mathrm{d}\phi \\
&= \int_{\phi=0}^{k\tau} \frac{\lambda^r \phi^{r-1} e^{-\lambda\phi}}{(r-1)!} \left[\sum_{j=0}^{s-r-1} \frac{\lambda^j (k\tau-\phi)^j}{j!} e^{-\lambda(k\tau-\phi)} \right] \mathrm{d}\phi \\
&= \sum_{j=0}^{s-r-1} \left[\frac{\lambda^{(j+r)}}{(j+r)!} e^{-\lambda k\tau} (k\tau)^{j+r-1} \times \cdots \right. \\
&\qquad \left. \times \int_{\phi=0}^{k\tau} \frac{(j+r)!}{(r-1)!j!} \left(1-\frac{\phi}{k\tau}\right)^j \left(\frac{\phi}{k\tau}\right)^{r-1} \mathrm{d}\phi \right].
\end{aligned}
$$

With the substitution $\varphi = \frac{\phi}{k\tau}$, the beta function integrates out:

$$
\begin{aligned}
A &= \sum_{j=0}^{s-r-1} \left[\frac{\lambda^{(j+r)}}{(j+r)!} e^{-\lambda k\tau} (k\tau)^{j+r-1} k\tau \int_{\varphi=0}^{1} \frac{(j+r)!}{(r-1)!j!} (1-\varphi)^j \varphi^{r-1} \mathrm{d}\varphi \right] \\
&= \sum_{j=0}^{s-r-1} \frac{(\lambda k\tau)^{j+r}}{(j+r)!} e^{-\lambda k\tau}.
\end{aligned}
$$

The same calculations can be used to derive the second part:

$$
B = \sum_{j=0}^{s-r-1} \frac{(\lambda k\tau)^{j+r}}{(j+r)!} e^{-\lambda k\tau} I_{1-\frac{1}{k}}(r, j+1).
$$

The difference $A - B$ results in Equation (5.7).

For a corrective repair, the state at the previous inspection was again less than r, but is greater than or equal to the failure state s at time $k\tau$. This probability is given by

$$
\Pr\{\text{CR in } ((k-1)\tau, k\tau]\} = \Pr\{S_r > (k-1)\tau, (k-1)\tau < S_s \leq k\tau\}
$$

and in this case becomes

$$
\begin{aligned}
&\Pr\{S_r > (k-1)\tau, (k-1)\tau < S_s \leq k\tau\} \\
&\quad = \Pr\{(k-1)\tau < S_r < S_s \leq k\tau\} \\
&\quad = \Pr\{(k-1)\tau < S_r \leq k\tau\} - \Pr\{(k-1)\tau < S_r \leq k\tau, S_s > k\tau\} \\
&\quad = \Pr\{S_r > (k-1)\tau\} - \Pr\{S_r > k\tau\} - \cdots \\
&\qquad - \Pr\{(k-1)\tau < S_r \leq k\tau, S_s > k\tau\} \\
&\quad = P(\lambda k\tau, r) - P(\lambda(k-1)\tau, r) - \cdots \\
&\qquad - \sum_{j=0}^{s-r-1} \frac{(\lambda k\tau)^{j+r}}{(j+r)!} e^{-\lambda k\tau} \left[1 - I_{1-\frac{1}{k}}(r, j+1)\right],
\end{aligned}
$$

where

$$
P(x, a) = \frac{1}{\Gamma(a)} \int_{t=0}^{x} t^{a-1} e^{-t} \mathrm{d}t
$$

is the incomplete gamma function, see also Abramowitz and Stegun (1965).

General formulation

The application of the model which was just described, is limited to Markov processes with successive phases as in Figure 5.1 and with identical exponential waiting times in each state. This is because the probabilities of a preventive and corrective repair have only been derived for this case. In this part, these probabilities are derived for any type of Markov process with successive phases as in Figure 5.1. These general results are obtained using a matrix notation and allow for state- and time-dependent transition intensities. If the transition intensities depend on the age of the process, the process is called instationary.

For any type of finite state Markov process, it is possible to calculate the transition probability function

$$
P_{ij}(s, t) = \Pr\{X(t) = j \mid X(s) = i\},
$$

which represents the probability of moving into state j after a duration t, given that the object was in state i at the beginning of this time period. The probability of being in state j after time t is simply determined by

$$
p_j(t) = \sum_{i=0}^{s} \Pr\{X(t) = j \mid X(0) = i\} \cdot \Pr\{X(0) = i\}.
$$

Using this probability, we can e.g. calculate the probability of not having reached state j by time t:

$$\Pr\{X(t) < j\} = \sum_{i=0}^{j-1} \Pr\{X(t) = i\} = \sum_{i=0}^{j-1} p_i(t).$$

To calculate the probability of a preventive or corrective repair, the current state of the deterioration process is used instead of the first passage time as in the special case described before. The probability of a preventive repair is the probability that the object was in a functional state at the previous inspection (i.e., $X((k-1)\tau) < r$) and in a marginal state at the current inspection (i.e., $r \leq X(k\tau) < s$):

$$\Pr\{\text{PR in } ((k-1)\tau, k\tau]\} = \Pr\{r \leq X(k\tau) < s, X((k-1)\tau) < r\}. \tag{5.8}$$

Similarly, the probability of corrective repair is the probability of the object being in a failed state at the current inspection when it was still in a functional state at the previous inspection:

$$\Pr\{\text{CR in } ((k-1)\tau, k\tau]\} = \Pr\{X(k\tau) \geq s, X((k-1)\tau) < r\}. \tag{5.9}$$

Equation (5.8) can be split up in two parts:

$$\begin{aligned}
&\Pr\{r \leq X(k\tau) < s, X((k-1)\tau) < r\} \\
&\quad = \Pr\{X(k\tau) < s, X((k-1)\tau) < r\} - \cdots \\
&\qquad - \Pr\{X(k\tau) < r, X((k-1)\tau) < r\}.
\end{aligned}$$

Denoting this difference as $A - B$, each probability can be calculated as follows:

$$\begin{aligned}
A &= \Pr\{X(k\tau) < s, X((k-1)\tau) < r\} \\
&= \Pr\{X(k\tau) < s \mid X((k-1)\tau) < r\}\ \Pr\{X((k-1)\tau) < r\} \\
&= \sum_{j=0}^{s-1} \Pr\{X(k\tau) = j \mid X((k-1)\tau) < r\}\ \Pr\{X((k-1)\tau) < r\} \\
&= \sum_{j=0}^{s-1} \sum_{i=0}^{r-1} \Pr\{X(k\tau) = j \mid X((k-1)\tau) = i\}\ \Pr\{X((k-1)\tau) = i\} \\
&= \sum_{j=0}^{s-1} \sum_{i=0}^{r-1} P_{ij}((k-1)\tau, k\tau) p_i((k-1)\tau)
\end{aligned}$$

and similarly

$$
\begin{aligned}
B &= \Pr\{X(k\tau) < r, X((k-1)\tau) < r\} \\
&= \Pr\{X(k\tau) < r \mid X((k-1)\tau) < r\} \cdot \Pr\{X((k-1)\tau) < r\} \\
&= \sum_{j=0}^{r-1} \Pr\{X(k\tau) = j \mid X((k-1)\tau) < r\} \cdot \Pr\{X((k-1)\tau) < r\} \\
&= \sum_{j=0}^{r-1} \sum_{i=0}^{r-1} P_{ij}((k-1)\tau, k\tau) p_i((k-1)\tau).
\end{aligned}
$$

Using these results, Equation (5.8) simplifies to

$$
A - B = \sum_{j=r}^{s-1} \sum_{i=0}^{r-1} P_{ij}((k-1)\tau, k\tau) p_i((k-1)\tau). \tag{5.10}
$$

Using the same approach and the knowledge that s is the final state, Equation (5.9) can be simplified as follows:

$$
\begin{aligned}
&\Pr\{X(k\tau) \geq s, X((k-1)\tau) < r\} \\
&\quad = \Pr\{X(k\tau) = s \mid X((k-1)\tau) < r\} \cdot \Pr\{X((k-1)\tau) < r\} \\
&\quad = \sum_{i=0}^{r-1} \Pr\{X(k\tau) = s \mid X((k-1)\tau) = i\} \cdot \Pr\{X((k-1)\tau) = i\} \\
&\quad = \sum_{i=0}^{r-1} P_{is}((k-1)\tau, k\tau) p_i((k-1)\tau). \qquad (5.11)
\end{aligned}
$$

Note that results (5.10) and (5.11) are valid for all types of Markov processes. For example, if we are using stationary Markov processes with time-independent transition rates, the transition functions $P_{ij}((k-1)\tau, \tau)$ reduce to $P_{ij}(\tau)$. These algorithmic formulations can therefore also be used to obtain the same results as are obtained using the analytical formulations in previous part.

The probability of a failure in the interval $(n-1, n]$ for n starting at the first year after the previous inspection, being $(k-1)\tau + 1$, to the time of the current inspection, being $k\tau$, is given by the following:

$$
\begin{aligned}
&\Pr\{X((k-1)\tau) < r, X(n-1) < s, X(n) = s\} \\
&\quad = \Pr\{X(n-1) < s, X(n) = s \,|\, X((k-1)\tau) < r\} \times \cdots \\
&\qquad \times \Pr\{X((k-1)\tau) < r\} \\
&\quad = \sum_{i=0}^{r-1} \Pr\{X(n-1) < s, X(n) = s \,|\, X((k-1)\tau) = i\} \times \cdots \\
&\qquad \times \Pr\{X((k-1)\tau) = i\} \\
&\quad = \sum_{i=0}^{r-1} \big(P_{is}(n) - P_{is}(n-1)\big) p_i((k-1)\tau).
\end{aligned}
$$

The probability of failure during each year in the interval $((k-1)\tau, k\tau]$ is therefore conditional on the probability that no preventive repair was performed at the start of this interval.

5.2.3 Application to Dutch bridge condition data

Two hypothetical scenarios are considered with fictitious costs presented in Table Table 5.1.

	Scenario A	Scenario B
Failure detection	immediate	by inspection
Inspection (c_I)	€1000	€1000
Prev. repair (c_P)	€10000	€10000
Corr. repair (c_F)	€40000	€10000
Unavailability (c_U)	N/A	€2000

Table 5.1: Costs for two hypothetical examples.

Scenario A considers the case in which a failure is immediately detected and repaired without the need for an inspection. The costs for a corrective repair are four times the costs of a preventive repair, therefore it will be economically interesting to repair before failure occurs. Scenario B considers the other case, in which failure is not detected until the next inspection and a cost is incurred for each unit of time that the superstructure is in a failed state. For both scenarios, the threshold for preventive repair is state $r = 3$ and the failed state is state $s = 5$. The unit time considered in both scenarios is one year.

Scenario B suits the case of superstructures very well, because state 5 is considered to be a condition failure and not an actual physical failure and therefore an inspection is required to assess the state of the superstructure.

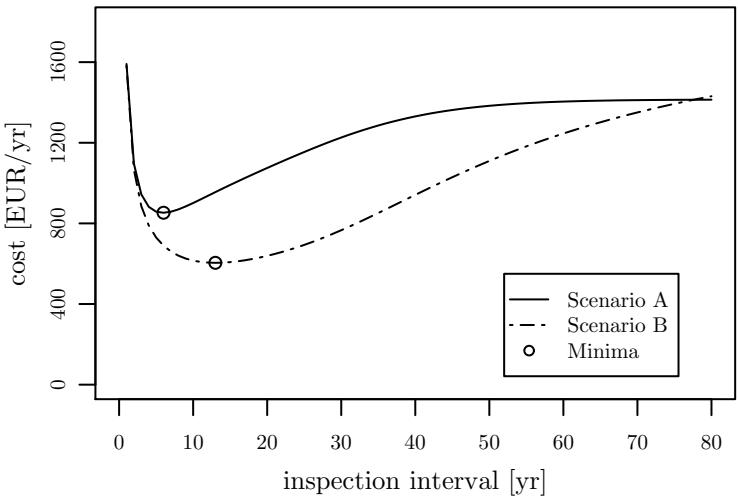


Figure 5.2: Expected average costs per year for superstructure maintenance as a function of the inspection interval τ and using model A.

For the bridge superstructures, the results for both scenarios are shown in Figure 5.2.

Since in scenario A corrective repair is expensive compared to preventive repair, the inspection interval with lowest expected average costs per year is shorter than the same optimal value for scenario B: 6 years compared to 13 years. With a mean time to preventive repair of $r/\lambda \approx 16.7$ years. This implies that about 3 inspections are performed for scenario A and a preventive repair is done after about 18 years. For scenario B, only about 2 inspections are performed and a repair is performed after about 26 years. As $\tau \to \infty$, the costs in scenario A converge to the costs of a corrective repair, €40000, divided by the expected lifetime of 28 years, which is approximately €1430. The costs in scenario B do not converge, but increase every year due to the cost of unavailability.

The cost model presented in this section is easier to implement than the classic policy improvement algorithm, which is used for Markov decision processes. Instead of presenting the decision maker with a single optimal value, this approach results in a clear graphical presentation as is demonstrated by Figure 5.2. Also, the models can be adjusted for various situations. For example, Equations (5.3) to (5.6) can be adjusted to include discounting, see e.g. van Noortwijk et al. (1997), or to include the time of the first inspection as an extra decision variable. The latter extension has been demonstrated before by Jia and Christer (2002) and is useful when the thresholds for preventive and corrective maintenance are very close to

each other. When the difference in costs for preventive or corrective replacement is large, this would result in a very short inspection interval, which would be unnecessary when the structure has just been taken into service. The model will determine the optimal combination of the time of first inspection and the subsequent periodic inspection interval.

Another extension is to include the threshold state for preventive maintenance as a decision variable. The optimal inspection intervals and the corresponding expected average costs per year for r ranging from state 1 to 4, are shown in Table 5.2.

Threshold	Example A		Example B	
r	τ	$\mathbb{E}C/\mathbb{E}I$	τ	$\mathbb{E}C/\mathbb{E}I$
	[yr]	[€/yr]	[yr]	[€/yr]
1	14	985	21	672
2	10	887	19	641
3	6	855	13	602
4	4	1026	9	592

Table 5.2: Optimal inspection intervals for different preventive maintenance threshold levels.

For example A, $r = 3$ is optimal and for example B $r = 4$ is optimal. In example B, the cost of a corrective repair is the same as the cost of a preventive repair, therefore the model minimizes the expected number of preventive renewals. This is achieved by placing the threshold for preventive repair as close as possible to the threshold for corrective repair, which is state 5 in this case. In general, the optimal length of the inspection interval τ becomes smaller as the threshold is moved closer to the threshold of a corrective renewal. As the margin for preventive maintenance becomes smaller, more frequent inspections are required to increase the probability of performing a preventive repair instead of a corrective repair. If $r = 4$ is chosen in example A, the risk of missing the opportunity for preventive repair becomes much greater and this results in a higher expectation for the average costs per year.

In the previous chapter, it was observed that a state-dependent model, namely model C, fits better to the bridge condition data compared to the state-independent model A, which is applied above. For superstructures, the estimated parameters for model C are listed in Table 4.5. Using the general formulation of the condition-based inspection model, the expected average costs per year are calculated using model C for superstructures. The result for both scenarios as defined in Table 5.1, is shown in Figure 5.3.

The optimal inspection interval for scenario A is 15 years with €654 per year and 23 years for scenario B with €498 per year. Because the mean

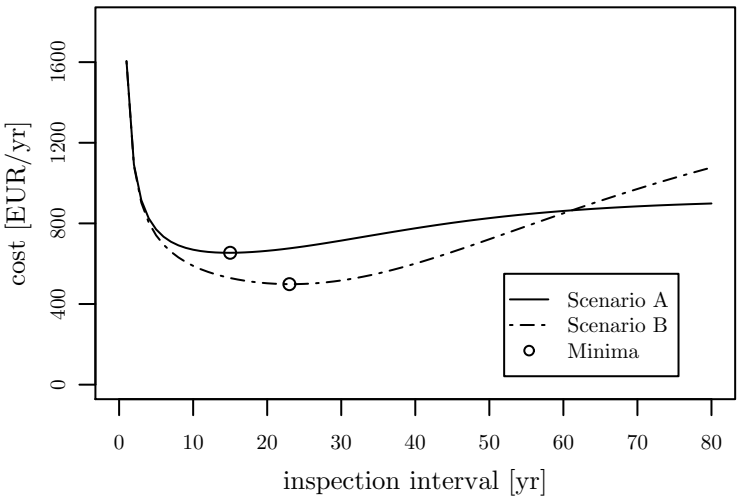


Figure 5.3: Expected average costs per year for superstructure maintenance as a function of the inspection interval τ and using model C.

times to reach the threshold states are longer in comparison with those given by model A, the times between inspections are also longer. However, the minima of both results are much less defined. For scenario A, the expected costs using an inspection interval of 10 or 20 years are only slightly more compared to the expected average costs per year for the optimal interval length of 15 years. This is a result of the much larger uncertainty in the amount of time it takes to reach the treshold states. In this case, the decision maker must make a decision with care. Using other criteria, like past experience or just common sense, he may decide for a shorter or longer period of time between inspections than the optimum given by the model. To aid in his decision, he may also calculate the variance of the results in Figures 5.2 and 5.3, which would give an indication about the confidence bounds around the mean. Explicit formulas for this purpose are given by van Noortwijk (2003) for the condition-based inspection model which includes discounting. Experience with the variance of the average costs per time unit, show that a choice for a shorter inspection interval generally has a smaller variance and is therefore less risky. So the decision maker might choose an inspection interval of 10 years for scenario A and about 20 years for scenario B.

5.3 SURVIVAL PROBABILITY

If no information on the cost of different maintenance actions is available, then a more simplistic approach to inspection planning could be used. A decision maker often needs to make a decision about whether or not a structure needs an earlier inspection than planned. For example, he may have a structure which is quite old, but has been reported to be in a good state. Not considering any life-cycle costs, he may want to assess the risk of a 'failure' before the next periodic inspection. For this purpose, the survival probability curves may be computed and reviewed by the decision maker. The selection of an inspection interval based on an acceptable risk level, is commonly used in risk-based inspection methodologies; see Kallen and van Noortwijk (2005a) for an introduction to risk-based inspection in the process and refining industry.

If T is the random variable representing the uncertain time to failure, the survival probability of a structure, given that it has survived up to age s, is defined as $\Pr\{T > t \,|\, T > s\}$. In other words, it is the probability of surviving up to age t conditional on having survived up to age s, where $0 \leq s \leq t$. A simple derivation shows that

$$\Pr\{T > t \,|\, T > s\} = \frac{\Pr\{T > t\}}{\Pr\{T > s\}} = \frac{S(t)}{S(s)},$$

where $S(t) = 1 - F(t) = \Pr\{T > t\}$ is the common notation for the survival probability function. For the finite-state Markov processes used here, the probability of failure is defined as the probability of reaching the final absorbing state, namely state 5. The 'survival' probability at age t is the probability of being in any state other than state 5 at age t: $S(t) \equiv \Pr\{X(t) \neq s\}$. For the sequential Markov process in Figure 5.1, this probability is

$$S(t) = \Pr\{X(t) < 5\} = \sum_{i=0}^{4} \Pr\{X(t) = i\} = \boldsymbol{p}_0' \exp\{\boldsymbol{R}t\}\mathbf{1},$$

where the matrix notation uses the fact that this probability distribution is a phase-type distribution as introduced in Section 2.1.3.

As an example, the survival probability curve of superstructures in state 0 at ages 0, 25, 50, 75, and 100, are calculated and shown in Figure 5.4. The survival curve for superstructures at age 0 is simply the random time required for a superstructure to reach state 5 from state 0, which has a mean of about 43 years as listed in Table 4.5. It is obvious that as superstructures reach higher ages, their remaining life shortens. However, the remaining life does not converge to zero, which is clearly observed in Table 5.3. This result is not entirely surprising, because it takes a minimum amount of

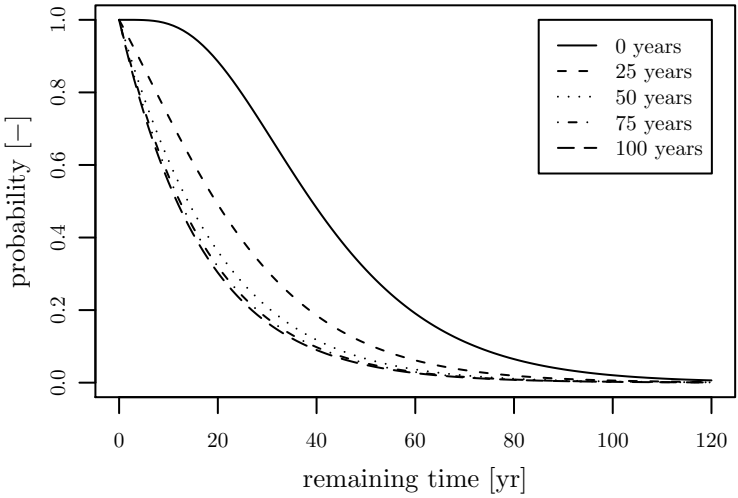


Figure 5.4: Survival probability curves for superstructures in the initial state 0 at various ages.

age [yr]	mean remaining time [yr]
0	42.7
25	24.7
50	19.5
75	17.8
100	17.1

Table 5.3: Mean time for superstructures to reach state 5 from state 0 at various ages.

time for a structure (or any object), which is in a perfect state at an old age, to reach the final state.

SUMMARY

Although both a Markov decision process and the model presented in Section 5.2 are condition-based decision models, they are different in many aspects. The most fundamental difference between them, is that the Markov decision processes are used to optimize the maintenance policy given a fixed inspection interval, whereas the condition-based model presented in Section 5.2 is used to optimize the length of the inspection interval given a fixed policy. The maintenance policy of the latter model is to perform a preventive repair or replacement if the structure is in a marginal state

and to perform a corrective repair or replacement if it is in a failed state. The length of time between inspections for the Markov decision processes described in Section 5.1 is implicitly included in the transition probability function $P_{ij}(a)$. If the policy of a Markov decision process is chosen to emulate the marginal checking policy of the model in Section 5.2, the time between inspections may be optimized by calculating the value function $V_\alpha(i)$ using different time periods for $P_{ij}(a)$.

A successful application of the condition-based maintenance model presented in this chapter depends on a number of aspects. First of all, there must be some relationship between the state in which the structure or component is and how much it costs to bring this object back to an as-good-as-new state. This is currently not possible for bridge condition data in the Netherlands, as there is no direct relationship between the condition state and the size and type of damages. This is due to the fact that discrete condition states are used for quantifying visual assessments of the overall condition of an object and not for sizing individual damages. The quote on page 23 formulates this rather well. This feature has significant consequences as it implies that practically all data in this form can not be used for an exact estimation of maintenance costs.

Another aspect which affects the applicability of a maintenance model, is the quality of the data. The aspect of data quality was discussed at the beginning of Chapter 4 and it is the registration of maintenance activities which is especially important in the context of maintenance optimization. If the data is of sufficient quality and the costs of maintenance are quantified, the condition-based model presented in Section 5.2 is easy to implement and easy to use. The implementation is straightforward compared to that of a Markov decision process and the results can be presented to decision makers in a simple plot like those in Figures 5.2 and 5.3.

Finally, it was shown in Section 5.3 how the survival curve of an object may be used to assess the risk of a 'failure' before the next inspection. The survival probability is calculated based on the current age and state of the object. This approach does not consider the costs of maintenance over the life-cycle of the object, but merely serves as an additional decision tool for the decision maker.

6

Conclusions and recommendations

In short, the research goal was to implement and test a suitable model for modeling the rate of deterioration for bridges in the Netherlands. There are over 3000 concrete bridges in the Netherlands, which are managed by the Civil Engineering Division of the Ministry of Transport, Public Works and Water Management. It is their duty to ensure the safety of the users and the availability of the road network by inspecting and maintaining these bridges. Due to an aging bridge stock and increasing demands on accountability, a structured approach to the planning and scheduling of maintenance activities is required. The primary uncertainty in the optimization of maintenance activities is the uncertainty in the rate of deterioration of structures. The approach presented in this research is aimed at the application of a model which uses the available condition data to quantify this uncertainty and therefore to aid in the decision making process.

SPECIFICATION

The approach advocated in this thesis, is of statistical nature and the first step in any statistical analysis is the specification of a model as a generator of the given data. As seen in Section 1.2.2, it is also possible to use a physics-based approach where the deterioration is explicitly modeled by a 'physical' model. The reason for choosing a statistical approach is the availability of a large data set with bridge condition states obtained by inspections over a period of almost twenty years. Also, physical measurements are infeasible at such a large scale.

To represent the uncertain development of the state of a structure, or one of its components over time, a finite-state Markov process (also known as a 'Markov chain') is used. This choice is quite common in civil engineering applications and also in medical studies, where the condition of a structure or patient is categorized according to a discrete and finite condition scale. For infrastructure in the Netherlands, a scale with seven states given in Table 1.1 is used. The use of a discrete condition scale is due to the visual nature of the inspections. The large number of structures in the network makes it too costly to perform actual damage size measurements on a regular basis. In the case of medical studies, there is often no device to measure the actual extent of a disease, therefore an expert assessment

is made using the information which is available. The reliability of a structure is traditionally used as a measure of the quality of a structure. This reliability depends on the ability of the structure to carry the loads that are placed upon it, which makes it a typical stress-strength model. There is no device to measure the reliability of a structure or component, therefore it must be calculated using the information available to the decision maker.

A Markov process is a stochastic process for which the Markov property holds. This property can be roughly defined as the property where the future state of the process is independent of the past given the present state. The assumption is therefore that all information from the past is contained in the current state of the process. In the past, there has been some debate within the bridge management community about the validity of this assumption. In statistical analysis, and especially in the context of maximum likelihood estimation, there is no 'true' model. The quality of a model is primarily determined by the likelihood of such a model generating the given data. Because all assumptions are made by choice, the decision maker may also use other subjective criteria to select a model for his purposes.

In Section 2.4, the testing of the Markov assumption is shortly addressed. For the available condition data in the Netherlands, obtained by periodic inspections, it is not practically feasible to test the validity of the Markov property in the data. The discrete nature of the data and the fact that such a large collection of data is available, weighs more heavily in the choice for a Markov process than the question of whether or not the Markov assumption is valid.

ESTIMATION

The Markov model used in this research is a parametric model, which means that there are a finite number of parameters. The Markov process itself is shaped by its transition structure, which is defined by the transition probability matrix or transition intensity matrix. Because a continuous-time Markov process is applied to the Dutch bridge condition data, a large part of the modeling aspect is the estimation of the transition intensities (or 'rates').

Chapter 3 reviews various aspects of the maximum likelihood estimation approach, which is used to fit the continuous-time Markov process to the panel data obtained by periodic visual inspections. These aspects include the consideration of uncertainty and variability in the inspection results and the testing of the influence of independent variables on the outcome. The maximum likelihood approach presented in Chapter 3 is different from previous approaches which are found in the literature and most of which are reviewed in Section 2.3. Contrary to many of the reviewed estimation

procedures, the maximum likelihood method is particularly well suited for panel data, because it does not depend on the exact times of transitions, nor on the exact duration of stay in any particular state. It only uses the likelihood of observing a structure in a particular state at a particular age, which is exactly the information available to the decision maker.

APPLICATION

Since 1985, a database to record inspections and maintenance activities on infrastructure in the Netherlands has been in use. In 2004, this database contained over 6000 observations of concrete bridge state conditions, which can be used to estimate the model parameters. The database was not specifically designed for use in a statistical analysis, therefore the data had to be prepared first. The process of preparing the data for application in a Markov model, consisted of filtering out faulty or incomplete data entries and converting the data set to a suitable format for the estimation procedure. Details on the data and the extraction process are given in Section 4.1. Experience with the Dutch database shows that the extraction process requires a great deal of attention and care in order to ensure the quality of the data set. Although this particular database is designed and used only by the Ministry of Transport, Public Works and Water Management in the Netherlands, it is very likely that databases in other countries or with a different application (like e.g., in sewer system and pavement management) will also require a significant effort in order to create a usable data set.

A Markov process is shaped by the transition probability or transition intensity matrix, therefore the layout of the transition structure is another decision for the modeler to make. The transition structure may be chosen such that only deterioration is possible (resulting in a 'progressive' process) or such that transitions towards better condition states are also allowed for. A progressive or sequential model should be used to model deterioration, since bridge conditions can not physically improve without intervention in the form of maintenance. Because of their rare application, self-healing materials are not considered here.

Age-dependent transition intensities

Existing Markov models in bridge management systems do not incorporate time-dependent transition probabilities or intensities. The approach presented here allows age-dependence to be included in a fairly straightforward manner. Unlike with semi-Markov processes, the dependence is not on the length of stay in a condition state, but rather on the age of the structure or the component. This allows the transition intensities to increase as the deteriorating object ages. It does not increase the probability of a pending

transition if the duration of stay in a condition state increases. As mentioned in Section 2.5, the latter feature results in a very complicated model which is not analytically tractable.

The addition of age-dependence results in a significantly better fit to the bridge condition data; see Section 4.2.1. However, this should not be considered as proof that bridge deterioration is age-dependent or non-homogeneous. It is only proof that the extra parameter helps to make the model fit closer to the observations. As for testing the Markov property, the condition data does not have sufficient detail for testing time-homogeneity.

Inspection variability

The inclusion of inspection variability is a topic which has received considerable interest, especially in combination with Markov decision processes. These approaches are referred to as 'partially observable' or 'latent' Markov decision processes. In the area of speech recognition, a Markov process with observational errors is called a hidden Markov model. The same terminology is used here, as the application of inspection variability is not combined with Markov decision processes. The primary goal in this research is to test if hidden Markov models are a useful extension to the regular Markov processes. The most challenging aspect of this model is to estimate the probability of misclassification. This is the probability of observing one of the condition states given a true underlying state. It may be estimated using the data itself or it may be determined by expert judgment.

In the end, hidden Markov models are not deemed to be suitable for application in bridge deterioration modeling. This is primarily due to the fact that, by the definition of visual inspections, the true state can never be measured and therefore not be used to check the quality of the estimated parameters. The course of the true state is completely determined by the choice for the misclassification probabilities and this choice is subjective.

Influence of independent variables

Although no two bridges will be exactly alike, most concrete bridges in the Netherlands can be considered to be similar in terms of design, construction material and usage. Nonetheless, it is possible to group bridges according to various characteristics. For example, bridges constructed using box girders may be considered separately from bridges with a beam supported road surface. It is also possible to group bridges according to their geographical location. Since the statistical analysis depends on the availability of a sufficient amount of data, it is important that the resulting groups do not contain a too small number of structures. In order to decide whether or not to consider certain groups of bridges, a statistical test of significance can be performed to determine if the groups behave significantly different.

If not, then it is not necessary for the decision maker to treat these groups differently in the maintenance model.

The maximum likelihood method is well suited to perform a so-called 'covariate analysis', where the influence of the independent variables (i.e., the variables which describe if a structure possesses a certain characteristic or not) on the outcome can be measured. From the Dutch database, three possible groups where identified in Section 4.4: 1) whether or not a bridge is located in the motorway, 2) whether the structure is identified as a 'bridge' or a 'viaduct', and 3) if the structure is located in a province with a relatively high population number. Only the last group behaves significantly different from structures located in provinces with a lower population number. When determining a maintenance policy, the decision maker may therefore include the province in which the structure is located as an additional variable.

Maintenance and inspection decisions

Chapter 5 is completely devoted to decision models which may be applied when using a finite-state Markov process as a deterioration model. The classical approach to optimizing maintenance decisions is to use a Markov decision process. An easy alternative condition-based maintenance model is presented in Section 5.2. This model balances the costs of inspections and preventive maintenance against the costs of corrective maintenance to determine the optimal time between inspections. It also allows the decision maker to determine the optimal threshold for preventive maintenance. Two scenarios have been evaluated in a case study using the deterioration model obtained in Chapter 4. In the first scenario, the bridge is found to be in an unacceptable state without the necessity of performing an inspection. The second scenario considers a more realistic situation where an inspection is required to observe if the structure has reached an unacceptable state or not. A penalty is applied for every year that a structure is left in an unacceptable state.

Because these decision models minimize the costs of inspections and maintenance over the lifetime of a structure, it is necessary to have representative cost data. Since the condition states used in bridge inspections do not represent a measure of the extent of damage, it is difficult to determine the cost of repairing structures in different states.

IMPLEMENTATION

In this project, special attention has been put into the aspects of the implementation of the maximum likelihood estimation and the maintenance model. Due to the large amount of state observations, it is not necessary to estimate the model parameters anew after every inspection. The model

is sufficiently robust such that significant changes in the parameter values are not to be expected. Nonetheless, a model for which the computations take too much time is not very appealing to the demanding user of today. Therefore, Chapter 7 deals with various techniques to calculate the transition probability function. This function gives the probability of transitioning between any two states during a specified period of time and if the transition intensities are age-dependent, this function will also depend on what age this period starts.

Out of all the reviewed approaches to calculating the transition probability function, uniformization is the most appealing. This particular approach enables an algorithmic calculation in which both the transition probability function and its derivatives may be calculated simultaneously. Also, other than scalar operations, the uniformization technique only consists of matrix multiplications and additions (or substractions), therefore avoiding relatively complicated calculations like the inversion of matrices. This feature makes it quite easy to implement the model in any programming language.

IN RETROSPECT

Several questions were defined in the introduction on page 12, which will be shortly addressed here:

a. it has indeed proven possible to extract condition data from the database in order to subject it to a statistical analysis.
b. most of the statistical models and estimation methods proposed for application to bridge maintenance, are reviewed in Chapter 2, but none possess all the properties that are desired for the Dutch situation.
c. a continuous-time Markov process is a suitable stochastic process for modeling the deterioration of structures. It is not as simplifying as a discrete-time Markov process and it not as intractable as a semi-Markov process with non-exponential waiting times. A likelihood function, which consists of a product of transition probabilities, has been defined. The method of maximum likelihood estimation is most suited for this non-linear model.
d. a bootstrapping procedure indicates that the model is robust to changes in the data. The estimates for the model parameters do not deviate very much when using different data sets obtained by bootstrapping the original data set.
e. the uniformization method is particularly efficient and well suited for calculating the transition probability function. Also, its relationship to the Poisson process enables the calculation of the transition probability function up to a specified accuracy.

f. the chosen model estimates that it takes about 45 years for a concrete bridge to reach state 5, which is a state at which major renovation is generally required. Due to the large variability in the many factors which influence the state of a structure, this estimate is very uncertain. With 90% probability, the time required to reach the fifth state is rougly between 14 and 96 years.
g. selecting different groups of bridges according to various characteristics in most cases did not result in significantly different estimates for the model parameters. Only a selection of structures according to the population density of the province in which they are located, resulted in significantly different transition intensities.
h. although it may seem interesting to explicitly model inspection variability in the model, the use of a hidden Markov model turned out to be practically infeasible. The primary problem is that the true state of the structure is never known, therefore it is quite difficult to determine the amount of misclassification by an inspector. Only expert judgment may be able to offer a quantification of the inspection errors.
i. for finite-state Markov processes, a suitable condition-based inspection and maintenance model has been defined. It determines the inspection interval with lowest expected average costs per time unit. A requirement for the application of such a model is that the costs of preventive and corrective repairs are available to the decision maker.

RECOMMENDATIONS

For the particular case of bridge management in the Netherlands, there are several recommendations for the application of the model proposed in this thesis. The first is to better record any maintenance activities in the database, such that backward transitions can more easily be related to either maintenance or inspection variability. This will allow to filter out maintenance data such that the deterioration model can be estimated using 'clean' condition data. Also, if maintenance activities can be filtered out, the application of a hidden Markov model for inspection variability may be possible. With the present data this did not result in satisfactory results. The application of such a model may also be made easier if inspectors document their reasoning behind the choice of a condition state. For example, if the inspector assigns a better condition to the structure compared to the previous inspection, he could indicate if this is because maintenance was performed after the previous inspection or not. In any case, this type of additional information should be given as a finite number of options by the system to the inspectors. It is not possible to automate the filtering of condition data if inspectors are allowed to supply their information in the form of unrestricted textual comments.

The database in the Netherlands contains a sufficient number of state observations for the application of the maximum likelihood estimation. The bootstrap procedure discussed in the section entitled "Distribution of the estimated parameters" on page 71, can be used to test if the estimator displays normality (i.e., if it is Gaussian). This test is particularly valuable for those states which are less frequently observed. However, the bootstrap is a computationally expensive procedure and it is important to ensure that the accurracy of the calculations is sufficiently high such that the outcome is not distorted. This research did not consider the situation where very little or no condition data is available. For those applications where this is the case, a Bayesian approach could be developed.

This research has shown that a finite-state Markov process is very well suited for use in a Dutch bridge management system. In order to complete the maintenance model with a decision model, it is necessary to develop a dependable cost model. This means that the costs of various maintenance actions must be quantified as a function of the existing condition state, the type of component or building material, and the age of the structure. This is a challenging problem because the condition rating scheme is specifically designed not to represent the actual size of damages. Past experience, combined with expert judgment, can be used to estimate maintenance costs.

The inspection intervals which are obtained using the condition-based maintenance model presented in Section 5.2, are often quite long. This is because the policy is assumed to start at the service start of the structure and structures deteriorate relatively slow. In most cases, structures are not new at the time of the analysis and have already undergone one or more inspections. Shorter and more reasonable inspection intervals may be obtained by adding the first inspection as a separate decision variable or by starting from the current state and age of the structure. The latter is mathematically possible due to the memoryless property of the continuous-time Markov process.

Finally, the problem at hand is a typical problem of decision making under uncertainty. Inspections are a tool to gain more information and therefore to reduce the uncertainty in the rate of deterioration of structures and their elements. However, it is not possible to acquire complete information and there is no true model to predict the lifetimes of structures. The best recommendation is therefore to use common sense when making any decision!

7

Appendix: transition probability function

Define the transition probability function as

$$p_{ij}(s,t) = \Pr\{X(t) = j \mid X(s) = i\}. \tag{7.1}$$

This function gives the probability of moving from state i at time $s \geq 0$ to state j a period $t - s \geq 0$ later. Let $\boldsymbol{P}(s,t) = \|p_{ij}(s,t)\|$ be the transition probability function matrix. $\boldsymbol{P}(s,t) = \boldsymbol{I}$ for all $s \geq 0$ if $t = s$ and $\boldsymbol{P}(s,t)$ is stochastic, that is: $\sum_j p_{ij}(s,t) = 1$ for all i. If the process is time-homogeneous, then $\boldsymbol{P}(s,t) = \boldsymbol{P}(0,t-s)$ with the equivalent shorthand notation defined as

$$p_{ij}(0,t-s) \equiv p_{ij}(t-s) = \Pr\{X(t-s) = j \mid X(0) = i\}. \tag{7.2}$$

This chapter deals with how to efficiently calculate the transition probability function and its sensitivity towards the model parameters. The parameters in the context of Markov processes are the transition intensities which, as is the case for inhomogeneous Markov processes, may be a function of the process age.

7.1 HOMOGENEOUS MARKOV PROCESSES

7.1.1 KOLMOGOROV EQUATIONS

The transition probability function is the solution to the forward and backward Kolmogorov equations. For a continuous-time Markov process with constant transition intensities, these are given by $\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{Q}$ and $\boldsymbol{P}'(t) = \boldsymbol{Q}\boldsymbol{P}(t)$ respectively. The well known solution to these differential equations is the matrix exponential as defined in Equation (2.6): $\boldsymbol{P}(t) = \exp\{\boldsymbol{Q}t\}$. There are a number of ways to show that this is the solution to the Kolmogorov equations. One approach is to use the transition probability function in Equation (2.4) and use a Laplace transformation to derive this result, see e.g., Howard (1971). The set of linear differential equations defined by the Kolmogorov equations can also be solved by the method of successive approximations. This is demonstrated next.

Take the backward Kolmogorov equations, then these define a boundary value problem of the form

$$\boldsymbol{X}' = \boldsymbol{Q}\boldsymbol{X}, \quad \boldsymbol{X}(0) = \boldsymbol{I},$$

where $\boldsymbol{I}$ is the identity matrix. Assume that a solution $\boldsymbol{X} = \boldsymbol{P}(t)$ exists, then the integral equation

$$\boldsymbol{P}(t) = \boldsymbol{I} + \int_{u=0}^{t} \boldsymbol{Q}\boldsymbol{P}(u)\,\mathrm{d}u \tag{7.3}$$

is equivalent to the boundary value problem in the sense that any solution of one is also a solution of the other. This can easily be verified by setting $t = 0$ to obtain the initial condition $\boldsymbol{P}(0) = \boldsymbol{I}$ and by taking the derivative towards t to obtain the set of differential differential equations $\boldsymbol{P}'(t) = \boldsymbol{Q}\boldsymbol{P}(t)$. The derivation can be done because the solution $\boldsymbol{P}(t)$ is assumed to be a continuous function and $\boldsymbol{Q}$ is constant over the range of integration. The initial approximation is the boundary condition $\boldsymbol{P}^{(0)}(t) = \boldsymbol{I}$. Substitution into Equation (7.3) yields the second approximation

$$\boldsymbol{P}^{(1)}(t) = \boldsymbol{I} + \int_{u=0}^{t} \boldsymbol{Q}\boldsymbol{P}^{(0)}(u)\,\mathrm{d}u = \boldsymbol{I} + \boldsymbol{Q}t + \boldsymbol{Q}^2\frac{t^2}{2}.$$

Through proof by induction, it is straightforward to show that

$$\boldsymbol{P}^{(n)}(t) = \boldsymbol{I} + \boldsymbol{Q}t + \boldsymbol{Q}^2\frac{t^2}{2} + \cdots + \boldsymbol{Q}^n\frac{t^n}{n!}$$

for $n \to \infty$. Then, by the definition of the matrix exponential in Equation (2.6), $\boldsymbol{P}^{(n)}(t) \to \exp\{\boldsymbol{Q}t\}$.

7.1.2 Matrix exponential

The definition of the matrix exponential in Equation (2.6) is restated here:

$$\exp\{\boldsymbol{Q}t\} = \boldsymbol{I} + \boldsymbol{Q}t + \boldsymbol{Q}^2\frac{t^2}{2!} + \boldsymbol{Q}^3\frac{t^3}{3!} + \cdots \tag{7.4}$$

Although the calculation of the matrix exponential may look trivial, it is certainly not. The formulation as an infinite series is an analytical result, but in practice it is impossible to calculate an infinite number of terms. A quasi analytical solution is to calculate the infinite series up to a large, but finite number of terms. According to Moler and van Loan (1978), the standard Taylor series approximation, achieved by truncating the infinite series in Equation (7.4) after a convergence criterion has been reached, is slow and prone to errors if a low floating point accuracy is used. The errors are due to rounding during the matrix multiplication and these in turn are a result of the presence of both positive and negative elements in the matrix $\boldsymbol{Q}$, see e.g., Ross (2000, p.350).

There are many ways to calculate the matrix exponential and Châtelet et al. (1996) grouped these in three categories:

1. numeric: integration methods for ordinary differential equations like e.g., the Euler and Runga-Kutta schemes,
2. analytical: e.g., using the Laplace transformation and the method based on matrix diagonalization,
3. quasi or partially analytical: e.g., exponential limit method and uniformization.

Matrix diagonalization

Assume that $\boldsymbol{A}$ is a square matrix of size $n \times n$ and has n linearly independent eigenvectors, then the matrix $\boldsymbol{A}$ is diagonalizable, i.e., it can be written in the form $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$, where $\boldsymbol{D}$ is a diagonal matrix. Matrix diagonalization is equivalent to finding the eigenvalues (and the eigenvectors) of $\boldsymbol{A}$, because the eigenvalues form the diagonal in $\boldsymbol{D}$. If $\boldsymbol{A}$ can be diagonalized, it holds that

$$\exp\{\boldsymbol{A}\} = \boldsymbol{P}\exp\{\boldsymbol{D}\}\boldsymbol{P}^{-1},$$

where $\exp\{\boldsymbol{D}\}$ is easily obtained by exponentiating the elements in the diagonal of $\boldsymbol{D}$. The method of matrix diagonalization for calculating the exponential of a square matrix is an analytical method. A sufficient condition for the matrix $\boldsymbol{A}$ to be diagonalizable, is that no two eigenvalues are the same (i.e., all n eigenvalues have multiplicity 1). By definition this excludes the possibility to model sequential continuous-time Markov processes with identical waiting times in each state. Other models may not have this problem, but numerical procedures as part of a maximum likelihood estimation may incidentally result in (nearly) identical eigenvalues.

Uniformization

Uniformization, also known as randomization, is a fairly well known way of computing the matrix exponential for Markov processes, see e.g., Tijms (2003) or Kohlas (1982). This approach is essentially an adapted Taylor series approximation. The uniformization technique makes the series approximation converge faster, avoids the rounding errors due to the negative elements in $\boldsymbol{Q}$, and also has a nice probabilistic interpretation.

Using the transition intensity matrix $\boldsymbol{Q}$ with elements defined in Equation (2.7), a new discrete-time Markov process $\overline{X}_t$ is generated by setting the one-step transition matrix to

$$\overline{p}_{ij} = \begin{cases} 1 - \lambda_i/\lambda, & \text{if } i = j, \\ (\lambda_i/\lambda)p_{ij}, & \text{if } i \neq j, \end{cases}$$

where $\lambda \geq \lambda_i$ for all i. If we now define a new continuous-time Markov process like $\overline{X}(t) = \overline{X}_{N(t)}$, then we get a Markov process where the transition times are generated by a Poisson process with rate $\lambda > 0$ and state transitions are generated by $\overline{p}_{ij}$. Whereas the process with intensity matrix $\boldsymbol{Q}$ (possibly) has a different rate for each state i, the new process has an intensity matrix $\overline{\boldsymbol{Q}}$ with the same rate λ for each state:

$$\overline{q}_{ij} = \begin{cases} -\lambda, & \text{if } i = j, \\ \lambda \overline{p}_{ij}, & \text{if } i \neq j. \end{cases}$$

If we use matrix notation, we say that $\overline{\boldsymbol{P}} = \boldsymbol{Q}/\lambda + \boldsymbol{I}$ and we can calculate the matrix exponential as

$$\boldsymbol{P}(0,t) = e^{\boldsymbol{Q}t} = e^{\lambda t(\overline{\boldsymbol{P}} - \boldsymbol{I})} = e^{\lambda t \overline{\boldsymbol{P}}} e^{-\lambda t \boldsymbol{I}} = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \overline{\boldsymbol{P}}^k. \tag{7.5}$$

For the fastest convergence of the series approximation, λ should be chosen as small as possible, but no smaller than the largest λ_i, otherwise $\overline{\boldsymbol{P}}$ will not be a stochastic matrix (i.e., not all rows sum to one). The best choice is therefore $\lambda = \max_i \lambda_i$.

When approximating Equation (7.5) by calculating the series up to $k = M$, the truncation error is given by the remainder of the series:

$$r_1(\lambda, t, M) = \sum_{k=M+1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \overline{\boldsymbol{P}}^k.$$

In order to determine how large M should be, it is necessary to assess the size of the truncation error. For this purpose, it is possible to calculate an upper bound for the error and this requires a matrix norm. Heidelberger and Goyal (1988) incorrectly use a matrix 1-norm, whereas the matrix infinity norm (maximum absolute row sum) defined as

$$\|\boldsymbol{A}\|_\infty = \max_i \sum_j |A_{ij}|, \tag{7.6}$$

is better suited, because $\|\overline{\boldsymbol{P}}^k\|_\infty = 1$ for all $k = 0, 1, 2, \ldots$. Using the triangular inequality with the matrix norm, the upper bound of the truncation error is

$$\|r_1(\lambda, t, M)\|_\infty \leq \sum_{k=M+1}^{\infty} \|\overline{\boldsymbol{P}}^k\|_\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} = 1 - \mathcal{P}_{\lambda t}(M),$$

where $\mathcal{P}_\phi(n)$ is the cumulative Poisson distribution with parameter $\phi \geq 0$. So using Equation (7.6), the error $r_1(\lambda, t, M)$ is conveniently bounded by

the area under the tail after M terms of the Poisson distribution with parameter λt.

7.2 NON-HOMOGENEOUS MARKOV PROCESSES

This section deals with an extension of the homogeneous Markov processes where time- or age-dependent transition intensities are allowed for. In the theory of linear systems, these are also referred to as 'time-variant' systems, because their primary driver (which are the transition intensities) vary over time. Yet another term is 'in- or non-stationary', which is commonly used for discrete-time Markov chains which have time varying transition probabilities. The terminology 'inhomogeneous' Markov processes is used here, due to their close relationship with the inhomogeneous Poisson process.

7.2.1 KOLMOGOROV EQUATIONS

Let $f_{i,t}(u)$, with $u \geq 0$, be the probability density function of the random waiting time in state i for a continuous-time Markov process (observed) at time t. From the time-invariant definition of the cumulative exponential distribution in Equation (2.5), this density may be defined as $f_{i,t}(u) = \lambda_i(t) \exp\{-\lambda_i(t)u\}$. From

$$\begin{aligned} p_{ij}(t, t+u) &= \Pr\{\text{transition from state } i \text{ to } j \text{ in } (t, t+u]\} \\ &= \Pr\{T_{i,t} \leq u, J_{N(t)+1} = j, J_{N(t)} = i\} \\ &= p_{ij} \Pr\{T_{i,t} \leq u \,|\, J_{N(t)+1} = j, J_{N(t)} = i\} \\ &= p_{ij} \int_{s=0}^{u} f_{i,t}(s) \, \mathrm{d}s \\ &= p_{ij} f_{i,t}(u)u + \mathcal{O}(u) \end{aligned}$$

and because $f_{i,t}(u) \approx \lambda_i(t)$ for a small u, it holds that $p_{ij}(t, t+u) = p_{ij}\lambda_i(t)u + \mathcal{O}(u)$. Using this result, the following lemma may be deduced:

$$\lim_{u \to 0} \frac{1 - p_{ii}(t, t+u)}{u} = \lambda_i(t)$$

and

$$\lim_{u \to 0} \frac{p_{ij}(t, t+u)}{u} = \lambda_i(t) p_{ij}.$$

In this section, the forward and backward Kolmogorov equations for time-variant or inhomogeneous Markov processes are derived. These are

$$\frac{\partial \boldsymbol{P}(s,t)}{\partial t} = \boldsymbol{P}(s,t)\boldsymbol{Q}(t) \tag{7.7}$$

and

$$\frac{\partial \boldsymbol{P}(s,t)}{\partial s} = -\boldsymbol{Q}(s)\boldsymbol{P}(s,t) \tag{7.8}$$

respectively. Either one of these two equations can be used to describe the dynamics in time of the Markov process. Given a transition rate matrix $\boldsymbol{Q}(t)$, the Kolmogorov equations give the solution to the transition probability function $\boldsymbol{P}(s,t)$. In other words: the characteristics of the Markov process are determined by the choice for $\boldsymbol{Q}(t)$ and $\boldsymbol{P}(s,t)$ is determined by solving one of the Kolmogorov equations Equation (7.7) and Equation (7.8). The Chapman-Kolmogorov equations are given by

$$p_{ij}(s,t) = \sum_k p_{ik}(s, s+u)p_{kj}(s+u,t), \tag{7.9}$$

where $s \leq s+u \leq t$. From Equation (7.9) we obtain

$$p_{ij}(s+u,t) - p_{ij}(s,t) = \sum_{forallk\neq i} -p_{ik}(s,s+u)p_{kj}(s+u,t) + \big[1 - p_{ii}(s,s+u)\big]p_{ij}(s+u,t).$$

Hence,

$$\lim_{u\to 0} \frac{p_{ij}(s+u,t) - p_{ij}(s,t)}{u} = \lim_{u\to 0}\left\{ \sum_{k\neq i} -\left[\frac{p_{ik}(s,s+u)}{u}\right] p_{kj}(s+u,t) + \cdots + \left[\frac{1-p_{ii}(s,s+u)}{u}\right] p_{ij}(s+u,t)\right\}.$$

The limit can be taken inside the summation (see for example Ross (1970)) to reveal the backward Kolmogorov equations:

$$\frac{\partial}{\partial s} p_{ij}(s,t) = \sum_{k\neq i} -\lambda_i(s)p_{ik}p_{kj}(s+u,t) + \lambda_i(s)p_{ij}(s+u,t). \tag{7.10}$$

Define

$$q_{ij}(t) = \begin{cases} -\lambda_i(t), & \text{if } i = j, \\ \lambda_i(t)p_{ij}, & \text{if } i \neq j. \end{cases} \tag{7.11}$$

Using $\boldsymbol{Q}(t) = \|q_{ij}(t)\|$, Equation (7.10) can be written in the matrix notation given by Equation (7.8).

For the forward Kolmogorov equation, a similar approach can be used to derive Equation (7.7). Using the Chapman-Kolmogorov equation again with $s \leq t-u \leq t$, the following result is obtained:

$$p_{ij}(s,t) - p_{ij}(s,t-u) = \sum_{k \neq j} p_{ik}(s,t-u)p_{kj}(t-u,t) - \big[1 - p_{jj}(t-u,t)\big]p_{ij}(s,t-u).$$

Dividing both sides by u and taking the limit $u \to 0$, a scalar representation of the forward Kolmogorov equation is obtained:

$$\frac{\partial}{\partial t} p_{ij}(s,t) = \sum_{k \neq j} p_{ik}(s,t-u)\lambda_k(t)p_{kj} - \lambda_j(t)p_{ij}(s,t). \tag{7.12}$$

Using the definition Equation (7.11) and a matrix notation, Equation (7.12) results in Equation (7.7).

It is tempting to write

$$\boldsymbol{P}(s,t) = \exp\left\{\int_{u=s}^{t} \boldsymbol{Q}(u)\ \mathrm{d}u\right\} \tag{7.13}$$

with

$$\exp\left\{\int_{u=s}^{t} \boldsymbol{Q}(u)\,\mathrm{d}u\right\} = \boldsymbol{I} + \int_{u=s}^{t} \boldsymbol{Q}(u)\ \mathrm{d}u + \frac{1}{2}\int_{u=s}^{t} \boldsymbol{Q}(u)\ \mathrm{d}u \int_{v=s}^{t} \boldsymbol{Q}(v)\ \mathrm{d}v + \cdots$$

as the solution for nonhomogeneous Markov processes, like Howard (1971) did on page 843. Unfortunately this is not a general solution, which can be easily shown by taking the derivative (Kailath (1980, Chapter 9)):

$$\begin{aligned}\frac{\partial}{\partial t}\exp\left\{\int_{u=s}^{t} \boldsymbol{Q}(u)\,\mathrm{d}u\right\} &= \\ \boldsymbol{Q}(t) + \frac{\boldsymbol{Q}(t)}{2}\int_{u=s}^{t} \boldsymbol{Q}(u)\,\mathrm{d}u + \int_{v=s}^{t} \boldsymbol{Q}(v)\,\mathrm{d}v\frac{\boldsymbol{Q}(t)}{2} + \cdots &\\ \neq \exp\left\{\int_{u=s}^{t} \boldsymbol{Q}(u)\,\mathrm{d}u\right\}\boldsymbol{Q}(t).&\end{aligned}$$

The solution in Equation (7.13) only holds if $\boldsymbol{Q}(t)$ and $\int \boldsymbol{Q}(u)du$ commute, which is not the case for most practical applications. The true general solution may be found in using a similar approach as with the homogeneous case in Section 7.1.1. Equivalent to Equation (7.3) for the time-homogeneous process, the solution of the forward Kolmogorov Equation (7.7) may be written in the form of an integral equation:

$$\boldsymbol{P}(s,t) = \boldsymbol{I} + \int_{u=s}^{t} \boldsymbol{P}(s,u)\boldsymbol{Q}(u)\,\mathrm{d}u. \tag{7.14}$$

The successive approximations are:

$$\boldsymbol{P}^{(0)}(s,t) = \boldsymbol{I},$$

$$\boldsymbol{P}^{(1)}(s,t) = \boldsymbol{I} + \int_{t_1=t_0}^{t} \boldsymbol{Q}(t_1)\,\mathrm{d}t_1,$$

$$\boldsymbol{P}^{(2)}(s,t) = \boldsymbol{I} + \int_{t_2=t_0}^{t} \boldsymbol{Q}(t_2)\,\mathrm{d}t_2 + \int_{t_2=t_0}^{t} \left\{ \int_{t_1=t_0}^{t_2} \boldsymbol{Q}(t_1)\,\mathrm{d}t_1 \right\} \boldsymbol{Q}(t_2)\,\mathrm{d}t_2$$

up to

$$\boldsymbol{P}^{(n)}(s,t) = \boldsymbol{I} + \int_{t_2=t_0}^{t} \boldsymbol{Q}(t_2)\,\mathrm{d}t_2 + \int_{t_2=t_0}^{t} \left\{ \int_{t_1=t_0}^{t_2} \boldsymbol{Q}(t_1)\,\mathrm{d}t_1 \right\} \boldsymbol{Q}(t_2)\,\mathrm{d}t_2,$$
$$\ldots + \int_{t_n=t_0}^{t} \int_{t_{n-1}=t_0}^{t_n} \cdots \int_{t_1=t_0}^{t_2} \boldsymbol{Q}(t_1)\boldsymbol{Q}(t_2)\cdots\boldsymbol{Q}(t_n)\,\mathrm{d}t_1\,\mathrm{d}t_2\cdots\mathrm{d}t_n.$$

where $s = t_0 < t_1 < \cdots < t_i < \cdots < t_n \leq t$. If $n \to \infty$, then $\boldsymbol{P}_n(s,t) \to \boldsymbol{P}(s,t)$ and the resulting series approximation

$$\boldsymbol{P}(s,t) = \boldsymbol{I} + \cdots \tag{7.15}$$
$$+ \sum_{n=1}^{\infty} \int_{t_n=t_0}^{t} \int_{t_{n-1}=t_0}^{t_n} \cdots \int_{t_1=t_0}^{t_2} \boldsymbol{Q}(t_1)\boldsymbol{Q}(t_1)\cdots\boldsymbol{Q}(t_n)\,\mathrm{d}t_1\,\mathrm{d}t_2\cdots\mathrm{d}t_n,$$

is known as the Peano-Baker series. Convergence of this successive approximation is proven by Dacunha (2005) for the standard problem $\boldsymbol{x}'(t) = \boldsymbol{A}(t)\boldsymbol{x}(t), \boldsymbol{x}(t_0) = \boldsymbol{x}_0$ in the theory of linear systems. It should be no surprise that, for a stationary process with $\boldsymbol{Q}(t) = \boldsymbol{Q}$, the Peano-Baker series solution in Equation (7.15) reduces to

$$\boldsymbol{P}(0,t) = \boldsymbol{I} + \boldsymbol{Q}t + \boldsymbol{Q}^2\frac{t^2}{2!} + \boldsymbol{Q}^3\frac{t^3}{3!} + \cdots,$$

which is, of course, the same as the result in Equation (2.6). This Taylor series solution has no intuitive (i.e., probabilistic) interpretation, therefore the Peano-Baker series solution in Equation (7.15) does not either. But, as in the stationary case, uniformization can be used to transform this solution to one with a probabilistic interpretation.

7.2.2 PRODUCT INTEGRATION

The formulation of the transition probability function in Equation (7.14) as an integral equation is merely an equivalent formulation of the solution to the forward Kolmogorov equations defined in Equation (7.7). The integral equation does not actually give the solution, it merely reformulates the

problem itself. In this section, another formulation is introduced: the product integral, which is denoted by

$$\boldsymbol{P}(s,t) = \prod_{(s,t]} (\boldsymbol{I} + \boldsymbol{Q}(\,\mathrm{d}\tau)), \tag{7.16}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{Q}(t)$ the transition intensity matrix as a function of time t. The notation $\prod$ was suggested by Gill and Johansen (1990) and is related to the product $\prod$ like the integral $\int$ is related to the sum $\sum$.

The formal definition of the product integral is the following:

$$\prod_{(0,t]} (1 - f(\,\mathrm{d}t)) = \lim_{\max_i |t_i - t_{i-1}| \to 0} \prod_i (1 - (f(t_i) - f(t_{i-1})))$$

where the limit is taken over a sequence of ever finer partitions of the interval $[0,t]$. The function f must be a real valued, cadlag function of bounded variation, defined on the finite interval $[0,t] \subset \mathbb{R}$. The definition may be extended to cases with more than two states. Let $\boldsymbol{F}$ be a $n \times n$ matrix valued function of bounded variation in which each component is right continuous with left hand limits. The product integral over the interval $[0,t]$ is now simply defined as

$$\prod_{(0,t]} (\boldsymbol{I} + \boldsymbol{F}(dt)) = \lim_{\max_i |t_i - t_{i-1}| \to 0} \prod_i (\boldsymbol{I} + \boldsymbol{F}(t_i) - \boldsymbol{F}(t_{i-1})).$$

As is proven in Gill and Johansen (1990), the transition probability matrix defined in Equation (7.16) using the product integral has the required properties, i.e.

- $\boldsymbol{P}(s,t)$ is a stochastic matrix, since each row sums to one,
- $\boldsymbol{P}(s,t) = \boldsymbol{P}(s,u)\boldsymbol{P}(u,t)$ for $0 \leq s \leq u \leq t \leq \infty$,
- $\boldsymbol{P}(s,s) = \boldsymbol{I}$ with $0 \leq s$ and
- $\boldsymbol{P}(s,t) \to \boldsymbol{I}$ as $t \downarrow s$.

Most importantly, it is shown by Gill and Johansen (1990) that the product integral defined in Equation (7.16) is the unique solution to the forward and backward Kolmogorov equations. The transition probability matrices for both the continuous-time Markov process (with constant transition rates) and the discrete-time Markov process (with transitions at fixed discrete time points) are special cases of $\boldsymbol{P}(s,t)$. This can be easily shown and is done in the next section.

Special cases

Consider the continuous-time Markov process with transition intensity matrix $\boldsymbol{Q}$. We want to derive that $\boldsymbol{P}(s,t) = \boldsymbol{P}(0,t-s) \equiv \boldsymbol{P}(t-s) =$

$\exp\{\boldsymbol{Q}(t-s)\}$, similar to Equation (2.6). If we define an equidistant partition on the interval $[s,t]$: $s = t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N = t$ with $t_i = s + i(t-s)/N$, then

$$\begin{aligned}
\boldsymbol{P}(s,t) &= \lim_{N\to\infty} \prod_{i=1}^{N} (1 + \boldsymbol{Q}(t_i) - \boldsymbol{Q}(t_{i-1})) \\
&= \lim_{N\to\infty} \prod_{i=1}^{N} \left[1 + \frac{\boldsymbol{Q}(t-s)}{N}\right] \\
&= \lim_{N\to\infty} \left[1 + \frac{\boldsymbol{Q}(t-s)}{N}\right]^N \qquad (7.17) \\
&= \exp\{\boldsymbol{Q}(t-s)\},
\end{aligned}$$

which is the well known solution for continuous-time Markov processes. If we assume that transitions can only occur at discrete points in time, we get a discrete-time Markov process. Assume that $s, t = 0, 1, 2, \ldots$ with $s < t$ and let $N = t - s$ such that the interval has partitions of size 1. Since N is fixed, Equation (7.17) reduces to $\boldsymbol{P}(s,t) = [\boldsymbol{I} + \boldsymbol{Q}]^{t-s}$. As

$$q_{ij} = \lim_{\mathrm{d}t \downarrow 1} \Pr\{X(\,\mathrm{d}t) = j \mid X(0) = i\} = \Pr\{X(1) = j \mid X(0) = i\} = p_{ij}$$

for $i \neq j$ and $q_{ii} = -\sum_{j \neq i} p_{ij}$, we get that $\boldsymbol{P}(s,t) = \boldsymbol{P}^{t-s}$. This is again the familiar solution for discrete-time Markov chains (with $\boldsymbol{P}$ being the one-step transition probability matrix) under the assumption that the process is stationary (or time-homogeneous).

7.2.3 EULER SCHEME

The transition probability function $\boldsymbol{P}(s,t)$ is the solution to the forward and backward Kolmogorov equations, which are differential equations for which many numerical approximation techniques exist. The Euler scheme is the most simple technique which approximates the derivative with a finite difference. For example, for the forward Kolmogorov Equation (7.7) this means that

$$\big[\boldsymbol{P}(s,t) - \boldsymbol{P}(s,t-\Delta t)\big]\Delta t^{-1} = \boldsymbol{P}(s,t-\Delta t)\boldsymbol{Q}(t),$$

which yields

$$\boldsymbol{P}(s,t) = \boldsymbol{P}(s,t-\Delta t)(\boldsymbol{Q}(t)\Delta t + \boldsymbol{I})$$

for $t \geq s + \Delta t$ and with a sufficiently small step size Δt. A standard procedure for the application of the Euler scheme is to partition the time

interval $[s,t]$ in N parts, such that $\Delta t = (t-s)/N$ and $t_k = s + \Delta t$ for $k = 0, 1, \ldots, N$. The transition probability function may then be iteratively approximated starting with $\boldsymbol{P}(s,s) = \boldsymbol{I}$ and continuing with $\boldsymbol{P}(s,t_k) = \boldsymbol{P}(s,t_{k-1})(\boldsymbol{Q}(t_k)\Delta t + \boldsymbol{I}), k = 1, 2, \ldots, N$.

The efficiency of this approach of course depends on Δt. The step size must be chosen sufficiently small to ensure adequate accuracy, but a very small step size will require a large number of matrix multiplications. Note the similarity of the Euler scheme with the approach used in the section titled "Special cases" on page 121 for the product integral. Approximating the product integral formulation of the transition probability function in Equation (7.16) with a large, but finite, number of products is therefore the same as approximating the Kolmogorov equations using the Euler scheme.

7.2.4 UNIFORMIZATION

Similar to van Moorsel and Wolter (1998), $\overline{\boldsymbol{P}}(t) = \boldsymbol{Q}(t)/\lambda + \boldsymbol{I}$ can be used to derive the uniformization equation for inhomogeneous Markov processes. Starting from the Peano-Baker series in Equation (7.15) this results in:

$$\boldsymbol{P}(s,t) = \tag{7.18}$$
$$\sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \int_{t_n=t_0}^{t} \int_{t_{n-1}=t_0}^{t_n} \cdots \int_{t_1=t_0}^{t_2} \overline{\boldsymbol{P}}(t_1) \cdots \overline{\boldsymbol{P}}(t_n) \frac{n!}{t^n} \, \mathrm{d}t_1 \cdots \mathrm{d}t_n,$$

where $\lambda > 0$ is now chosen such that

$$\lambda \geq \inf_{t \in (s,t]} |Q_{ij}(t)|.$$

Equation (7.18) represents a inhomogeneous Markov process where the transition times occur according to a Poisson process with rate $\lambda > 0$ and transitions in the embedded Markov chain are performed according to $\overline{\boldsymbol{P}}(t)$. The product $\overline{\boldsymbol{P}}(t_1)\overline{\boldsymbol{P}}(t_2) \cdots \overline{\boldsymbol{P}}(t_n)$ is the n-step probability of transitions occurring at times $\{t_1, t_2, \ldots, t_n\}$. The integration is done over all possible sets $\{t_1, t_2, \ldots, t_n\}$ and $n!/t^n$ is the density of the order statistic of an n-dimensional uniform distribution resulting from the Poisson process. The paper by van Moorsel and Wolter (1998) continues with an algorithm to compute Equation (7.18), but it is clear that this is not a simple thing to do.

There is one way to reduce the complexity of Equation (7.18) and that is to choose the transition intensity matrix such that it can be decomposed like $\boldsymbol{Q}(t) = f(t)\boldsymbol{Q}$. In this way, we end up with a transition probability matrix for the embedded Markov chain which does not depend on time, i.e., $\overline{\boldsymbol{P}} = \boldsymbol{Q}(t)/f(t) + \boldsymbol{I} = \boldsymbol{Q} + \boldsymbol{I}$. As was pointed out by Rindos et al. (1995), Equation (7.13) is now a correct solution and it is possible to write

$$\boldsymbol{P}(0,t) = \exp\{\boldsymbol{Q}\Lambda(t)\} = \exp\{(\boldsymbol{Q}+\boldsymbol{I})\Lambda(t) - \boldsymbol{I}\Lambda(t)\} = \sum_{n=0}^{\infty} \overline{\boldsymbol{P}}^n \frac{\Lambda(t)^n}{n!} e^{-\Lambda(t)},$$

or, more general,

$$\boldsymbol{P}(s,t) = \sum_{n=0}^{\infty} \overline{\boldsymbol{P}}^n \frac{[\Lambda(t)-\Lambda(s)]^n}{n!} e^{-[\Lambda(t)-\Lambda(s)]}, \tag{7.19}$$

where $\Lambda(t) = \int_{u=0}^{t} f(u)\,\mathrm{d}u$. Now we have a inhomogeneous Poisson process generating the transition times and $\overline{\boldsymbol{P}}$ defining the transitions of the embedded Markov chain. This approach can only be used if the transition intensity matrix is of the form $\boldsymbol{Q}(t) = \boldsymbol{Q}f(t)$, where $\boldsymbol{Q}$ is a time-invariant (or constant) matrix and $f(t) \geq 0$ is a scalar function of t. An example of a matrix which is of this form is

$$\boldsymbol{Q}(t) = \begin{bmatrix} -a_1 t^b & a_1 t^b & 0 & 0 & 0 & 0 \\ 0 & -a_2 t^b & a_2 t^b & 0 & 0 & 0 \\ 0 & 0 & -a_3 t^b & a_3 t^b & 0 & 0 \\ 0 & 0 & 0 & -a_4 t^b & a_4 t^b & 0 \\ 0 & 0 & 0 & 0 & -a_5 t^b & a_5 t^b \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where $f(t) = t^b$.

7.3 PARAMETER SENSITIVITY

In order to maximize the likelihood function, which includes the matrix exponential as the transition probability function, it is necessary to assess the sensitivity of the model towards changes in the parameter values. For this, it is required to calculate the derivative of the matrix exponential towards each of the parameters contained in the transition matrix $\boldsymbol{Q}$. These derivatives can then be used to implement an iterative approximating scheme to find the values of the parameters which maximize the likelihood function. If $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_n\}$ is a vector of all parameters contained in $\boldsymbol{Q}(t)$, the partial derivative towards one of these parameters is given by

$$\frac{\partial}{\partial \lambda_i}\boldsymbol{P}(s,t) = \int_{u=s}^{t} \boldsymbol{P}(s,u)\frac{\partial \boldsymbol{Q}}{\partial \lambda_i}\boldsymbol{P}(u,t)\,\mathrm{d}u. \tag{7.20}$$

For a stationary Markov process with constant $\boldsymbol{Q}$, this becomes

$$\frac{\partial}{\partial \lambda_i}\exp\{\boldsymbol{Q}t\} = \int_{u=0}^{t} e^{(t-u)\boldsymbol{Q}}\frac{\partial \boldsymbol{Q}}{\partial \lambda_i}e^{u\boldsymbol{Q}}\,\mathrm{d}u, \tag{7.21}$$

which according to Tsai and Chan (2003) is due to Wilcox (1967). It is interesting to see the similarity with the case of a Markov chain. If we

assume $\boldsymbol{P}$ is a one-step transition probability matrix, then the derivative of the n-step probability towards the parameter λ_i is

$$\frac{\partial \boldsymbol{P}^n}{\partial \lambda_i} = \sum_{k=0}^{n-1} \boldsymbol{P}^k \left(\frac{\partial \boldsymbol{P}}{\partial \lambda_i} \right) \boldsymbol{P}^{n-1-k}.$$

The fact that Equation (7.21) holds for the derivative and not the well known result for the scalar case, can easily be shown as follows:

$$\begin{aligned} \frac{\partial}{\partial \lambda_i} \exp\{\boldsymbol{Q}t\} = \sum_{k=0}^{\infty} \frac{\partial \boldsymbol{Q}^k}{\partial \lambda_i} \cdot \frac{t^t}{k!} &= \sum_{k=0}^{\infty} \left[\sum_{i=0}^{n-1} \boldsymbol{Q}^i \left(\frac{\partial \boldsymbol{Q}}{\partial \lambda_i} \right) \boldsymbol{Q}^{k-1-i} \right] \frac{t^k}{k!} \\ &\neq \exp\{\boldsymbol{Q}t\} \cdot \frac{\partial \boldsymbol{Q}}{\partial \lambda_i}, \end{aligned}$$

where the partial derivative of $\boldsymbol{Q}$ towards λ_i can not be taken outside the summation due to the fact that matrix multiplication is not commutative (i.e., the order of multiplication of matrices can generally not be changed: $AB \neq BA$).

7.3.1 Matrix diagonalization

Here, the exact analytical derivative of $\boldsymbol{P}(s,t)$ towards the parameters in $\boldsymbol{Q}(t)$ will be derived under the assumption that $\boldsymbol{Q}(t)$ can be decomposed into $f(t)\boldsymbol{Q}$ as in the previous section.

The approach is based on the principle that the intensity matrix may be decomposed as $\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^{-1}$, where $\boldsymbol{D}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{Q}$ on the diagonal and the columns of $\boldsymbol{A}$ are the eigenvectors of $\boldsymbol{Q}$. Using diagonalization, the matrix exponential can be simplified as follows:

$$\boldsymbol{P}(s,t) = \exp\left\{ \boldsymbol{Q} \int_{u=s}^{t} f(u)\, \mathrm{d}u \right\} = \boldsymbol{A} \exp\{\boldsymbol{D}[\Lambda(t) - \Lambda(s)]\} \boldsymbol{A}^{-1}.$$

For ease of notation, let $\boldsymbol{E} = \exp\{\boldsymbol{D}[\Lambda(t) - \Lambda(s)]\}$ with elements $e_i = \exp\{d_i[\Lambda(t) - \Lambda(s)]\}, i = 1, \ldots, n$ on the diagonal, where d_i is the i-th diagonal element of matrix $\boldsymbol{D}$. In Kalbfleisch and Lawless (1985), the derivation for $\partial \boldsymbol{P}/\partial \theta_i$ is included, but the following essential steps, taken from Jennrich and Bright (1976), are not mentioned in the derivation. Using the product rule for differentiation and a shorthand notation $d\boldsymbol{P}(s,t)$ for the partial derivative of the transition probability function towards one of the parameters, the following is obtained:

$$\mathrm{d}\boldsymbol{P}(s,t) = (\,\mathrm{d}\boldsymbol{A})\boldsymbol{E}\boldsymbol{A}^{-1} + \boldsymbol{A}(\,\mathrm{d}\boldsymbol{E})\boldsymbol{A}^{-1} + \boldsymbol{A}E(\,\mathrm{d}\boldsymbol{A}^{-1}),$$

such that

$$\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{P}(s,t)\boldsymbol{A} = \boldsymbol{A}^{-1}(\,\mathrm{d}\boldsymbol{A})\boldsymbol{E} + \,\mathrm{d}\boldsymbol{E} + \boldsymbol{E}(\,\mathrm{d}\boldsymbol{A}^{-1})\boldsymbol{A}. \tag{7.22}$$

To obtain the ij-th element, the following can be done:

$$\begin{aligned}(\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{P}(s,t)\boldsymbol{A})_{ij} &= \sum_k (\boldsymbol{A}^{-1}\mathrm{d}\boldsymbol{A})_{ik}(\boldsymbol{E})_{kj} + (\mathrm{d}\boldsymbol{E})_{ij} + \sum_k (\boldsymbol{E})_{ik}(\mathrm{d}\boldsymbol{A}^{-1}\boldsymbol{A})_{kj} \\ &= (\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A})_{ij}e_j + (\mathrm{d}\boldsymbol{E})_{ij} + e_i(\mathrm{d}\boldsymbol{A}^{-1}\boldsymbol{A})_{ij},\end{aligned}$$

where the fact that $(\boldsymbol{E})_{ij} = e_j$ if $i = j$ is used and $(\boldsymbol{E})_{ij} = 0$ otherwise. Using the following from Jennrich and Bright (1976):

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \Rightarrow \,\mathrm{d}\boldsymbol{A}^{-1}\boldsymbol{A} + \boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A} = 0 \Rightarrow \,\mathrm{d}\boldsymbol{A}^{-1}\boldsymbol{A} = -\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A} \tag{7.23}$$

to obtain

$$(\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{P}(s,t)\boldsymbol{A})_{ij} = \begin{cases} (e_j - e_i)(\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A})_{ij} & \text{if } i \neq j, \\ \mathrm{d}e_i & \text{if } i = j. \end{cases} \tag{7.24}$$

Now $\mathrm{d}\boldsymbol{A}$ remains, which can be removed by using a similar combination of Equation (7.22) and Equation (7.23) for $\mathrm{d}\boldsymbol{Q}$. Namely, by taking $\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{Q}\boldsymbol{A} = \boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A}\boldsymbol{D} + \,\mathrm{d}\boldsymbol{D} - \boldsymbol{D}A^{-1}\,\mathrm{d}\boldsymbol{A}$. If the non-diagonal elements are equated, this results in

$$(\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{A})_{ij} = \frac{(\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{Q}\boldsymbol{A})_{ij}}{d_j - d_i} \quad \text{for } i \neq j,$$

which can be substituted in Equation (7.24) in order to get

$$(\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{P}(s,t)\boldsymbol{A})_{ij} = \begin{cases} \frac{e_j - e_i}{d_j - d_i}(\boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{Q}\boldsymbol{A})_{ij} & \text{if } i \neq j, \\ de_i & \text{if } i = j. \end{cases}$$

If we let $\boldsymbol{V} = \boldsymbol{A}^{-1}\,\mathrm{d}\boldsymbol{P}(s,t)\boldsymbol{A}$ as in Kalbfleisch and Lawless (1985), we get that

$$\mathrm{d}\boldsymbol{P}(s,t) = \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^{-1}. \tag{7.25}$$

The diagonalization approach gives an exact result in the form of Equation (7.25) to calculate the derivative of the transition probability function. This allows for the efficient calculation of the sensitivity of the model towards changes in the model parameters. However, this approach assumes that the matrix $\boldsymbol{Q}$ can be diagonalized, which implicitly assumes that the necessary computational tools to perform the diagonalization are available and that the matrix $\boldsymbol{Q}$ is non-singular (i.e., invertible). Problems arise when $\boldsymbol{Q}$ is close to being singular or when $d_j = d_i$ for $i \neq j$. Even if the initial situation is well formed, there is little control over the parameters of the model when using this approach in a Newton-Raphson iterative scheme.

Thus the iteration may halt unexpectedly due to the aforementioned problems.

7.3.2 UNIFORMIZATION

Homogeneous Markov processes

Using the same approach as Heidelberger and Goyal (1988), Equation (7.5) can be differentiated to λ_i:

$$\frac{\partial}{\partial \lambda_i} \boldsymbol{P}(0,t) = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \frac{\partial}{\partial \lambda_i} \overline{\boldsymbol{P}}^k \tag{7.26}$$

and, because

$$\frac{\partial}{\partial \lambda_i} \overline{\boldsymbol{P}}^k = \overline{\boldsymbol{P}}^{k-1} \cdot \frac{\partial}{\partial \lambda_i} \overline{\boldsymbol{P}} + \frac{\partial}{\partial \lambda_i} \overline{\boldsymbol{P}}^{k-1} \cdot \overline{\boldsymbol{P}},$$

the derivative of $\overline{\boldsymbol{P}}^k$ can be computed iteratively using the values $\overline{\boldsymbol{P}}^{k-1}$ and its derivative from the previous step. The matrices $\partial \overline{\boldsymbol{P}} / \partial \lambda_i = \lambda^{-1} \partial \boldsymbol{Q} / \partial \lambda_i$ and $\overline{\boldsymbol{P}}$ only have to be computed once at the start of the iteration scheme. Note that when we take $\lambda = \max_i \lambda_i$, we assume λ is just a constant when taking the derivative to λ_i.

Like with the matrix exponential, it is also possible to determine the bound on the error if we calculate the derivative in Equation (7.26) up to $k = M$. Let

$$r_2(\lambda, t, M) = \sum_{k=M+1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \frac{\partial}{\partial \lambda_i} \overline{\boldsymbol{P}}^k.$$

It is easy to show that

$$\left\| \frac{\partial}{\partial \lambda_i} \overline{\boldsymbol{P}}^k \right\|_\infty \leq \frac{k}{\lambda} \left\| \frac{\partial \boldsymbol{Q}}{\partial \lambda_i} \right\|_\infty,$$

such that

$$\begin{aligned} \|r_2(\lambda, t, M)\|_\infty &\leq \sum_{k=M+1}^{\infty} \frac{k}{\lambda} \left\| \frac{\partial \boldsymbol{Q}}{\partial \lambda_i} \right\|_\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} \\ &= \frac{(\lambda t)}{\lambda} \left\| \frac{\partial \boldsymbol{Q}}{\partial \lambda_i} \right\|_\infty \sum_{k=M+1}^{\infty} e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} \end{aligned}$$

and therefore the error is again conveniently bounded by

$$\|r_2(\lambda, t, M)\|_\infty \leq \left\| \frac{\partial \boldsymbol{Q}}{\partial \lambda_i} \right\|_\infty t\big[1 - \mathcal{P}_{\lambda t}(M-1)\big].$$

The second derivative may be calculated in a similar fashion. For any two parameters λ_i and λ_j, the derivative of $\overline{\boldsymbol{P}}^k$ can be calculated recursively as follows:

$$\frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}}^k = \overline{\boldsymbol{P}}^{k-1}\cdot\frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}} + \frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}}^{k-1}\cdot\overline{\boldsymbol{P}},$$

and

$$\frac{\partial}{\partial\lambda_j}\overline{\boldsymbol{P}}^k = \overline{\boldsymbol{P}}^{k-1}\cdot\frac{\partial}{\partial\lambda_j}\overline{\boldsymbol{P}} + \frac{\partial}{\partial\lambda_j}\overline{\boldsymbol{P}}^{k-1}\cdot\overline{\boldsymbol{P}}.$$

The second derivative can also easily be calculated recursively:

$$\begin{aligned}\frac{\partial^2}{\partial\lambda_j\partial\lambda_j}\overline{\boldsymbol{P}}^k = \frac{\partial}{\partial\lambda_j}\overline{\boldsymbol{P}}^{k-1}\cdot\frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}} + \overline{\boldsymbol{P}}^{k-1}\cdot\frac{\partial^2}{\partial\lambda_j\partial\lambda_i}\overline{\boldsymbol{P}} + \ldots \\ \ldots + \frac{\partial^2}{\partial\lambda_j\partial\lambda_i}\overline{\boldsymbol{P}}^{k-1}\cdot\overline{\boldsymbol{P}} + \frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}}^{k-1}\cdot\frac{\partial}{\partial\lambda_j}\overline{\boldsymbol{P}}\end{aligned}$$

and because $\partial^2\overline{\boldsymbol{P}}/\partial\lambda_j\partial\lambda_i = \mathbf{0}$, this reduces to

$$\frac{\partial^2}{\partial\lambda_j\partial\lambda_j}\overline{\boldsymbol{P}}^k = \frac{\partial}{\partial\lambda_j}\overline{\boldsymbol{P}}^{k-1}\cdot\frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}} + \frac{\partial^2}{\partial\lambda_j\partial\lambda_i}\overline{\boldsymbol{P}}^{k-1}\cdot\overline{\boldsymbol{P}} + \frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}}^{k-1}\cdot\frac{\partial}{\partial\lambda_j}\overline{\boldsymbol{P}}.$$

Each element in this last equation is calculated in a previous step, therefore no extra effort (other than a simple addition) is required to calculate the second derivative. As

$$\|\overline{\boldsymbol{P}}\|_\infty = 1 \quad\text{and}\quad \left\|\frac{\partial}{\partial\lambda_i}\overline{\boldsymbol{P}}^k\right\|_\infty \le \frac{k}{\lambda}\left\|\frac{\partial\boldsymbol{Q}}{\partial\lambda_i}\right\|_\infty,$$

it is possible to determine an upper bound to the second derivative of the matrix $\overline{\boldsymbol{P}}^k$:

$$\begin{aligned}\left\|\frac{\partial^2}{\partial\lambda_j\partial\lambda_j}\overline{\boldsymbol{P}}^k\right\|_\infty &\le \frac{2(k-1)}{\lambda^2}\left\|\frac{\partial\boldsymbol{Q}}{\partial\lambda_i}\right\|_\infty\cdot\left\|\frac{\partial\boldsymbol{Q}}{\partial\lambda_j}\right\|_\infty + \left\|\frac{\partial^2}{\partial\lambda_j\partial\lambda_i}\overline{\boldsymbol{P}}^{k-1}\right\|_\infty \\ &\le \left\|\frac{\partial\boldsymbol{Q}}{\partial\lambda_i}\right\|_\infty\cdot\left\|\frac{\partial\boldsymbol{Q}}{\partial\lambda_j}\right\|_\infty\cdot\left\{\frac{2(k-1)}{\lambda^2} + \frac{2(k-2)}{\lambda^2} + \ldots + \frac{2}{\lambda^2}\right\}\end{aligned}$$

Since

$$\left\{\frac{2(k-1)}{\lambda^2} + \frac{2(k-2)}{\lambda^2} + \ldots + \frac{2}{\lambda^2}\right\} = \frac{2}{\lambda^2}\sum_{i=1}^{k-1} i = \frac{2}{\lambda^2}\left\{\frac{k(k-1)}{2}\right\},$$

the final result can be reduced to

$$\left\| \frac{\partial^2}{\partial\lambda_j\partial\lambda_j}\overline{\boldsymbol{P}}^k \right\|_\infty \leq \left\| \frac{\partial \boldsymbol{Q}}{\partial\lambda_i} \right\|_\infty \cdot \left\| \frac{\partial \boldsymbol{Q}}{\partial\lambda_j} \right\|_\infty \cdot \frac{k(k-1)}{\lambda^2}$$

Now because

$$\left\| \frac{\partial \boldsymbol{Q}}{\partial\lambda_i} \right\|_\infty = \begin{cases} 2, & \text{if } \lambda_i, \text{ for all } i, \text{ is in } \boldsymbol{Q} \text{ or} \\ 0, & \text{otherwise,} \end{cases}$$

it holds that

$$\left\| \frac{\partial^2}{\partial\lambda_j\partial\lambda_j}\overline{\boldsymbol{P}}^k \right\|_\infty \leq \begin{cases} k(k-1)(\frac{2}{\lambda})^2, & \lambda_i \text{ and } \lambda_j \text{ are in } \boldsymbol{Q} \text{ or,} \\ 0, & \text{one of } \lambda_i \text{ or } \lambda_j \text{ is not in } \boldsymbol{Q}. \end{cases}$$

To calculate the second derivative of the transition probability function

$$\frac{\partial^2}{\partial\lambda_j\partial\lambda_i}\boldsymbol{P}(0,t) = \sum_{k=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\frac{\partial^2}{\partial\lambda_j\partial\lambda_i}\overline{\boldsymbol{P}}^k$$

up to $k = M$, we can determine an upper bound on the error defined by the remainder of the infinite series. Let

$$\|r_3(\lambda, t, M)\|_\infty = \sum_{k=M+1}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\frac{\partial^2}{\partial\lambda_j\partial\lambda_i}\overline{\boldsymbol{P}}^k,$$

then the bound is given by

$$\begin{aligned} \|r_3(\lambda, t, M)\|_\infty &\leq \sum_{k=M+1}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\left\| \frac{\partial^2}{\partial\lambda_j\partial\lambda_j}\overline{\boldsymbol{P}}^k \right\|_\infty \\ &\leq \sum_{k=M+1}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}k(k-1)\left(\frac{2}{\lambda}\right)^2 = \sum_{k=M+1}^{\infty} e^{-\lambda t}\frac{(\lambda t)^{k-2}}{(k-2)!}(2t)^2. \end{aligned}$$

If $\mathcal{P}_\phi(n)$ is the cumulative Poisson distribution with parameter $\phi \geq 0$, then the upper bound can be written as

$$\|r_3(\lambda, t, M)\|_\infty \leq (2t)^2[1 - \mathcal{P}_{\lambda t}(M-2)].$$

Non-homogeneous Markov processes

From a practical point of view, it is not feasible to calculate the derivative of the time-dependent transition probability function $\boldsymbol{P}(s,t)$ using the formulation given by Equation (7.18). It is however quite possible to do this for the randomized version in Equation (7.19), which was derived under the assumption that the transition intensity could be decomposed as $\boldsymbol{Q}(t) = \boldsymbol{Q}f(t)$. Let

$$g_k(s,t) = \frac{[\Lambda(t) - \Lambda(s)]^k}{k!}e^{-[\Lambda(t)-\Lambda(s)]},$$

then the derivative of the transition probability function towards the parameter ν_i is determined by

$$\frac{\partial}{\partial \nu_i}\boldsymbol{P}(s,t) = \sum_{k=0}^{\infty}\left[\frac{\partial}{\partial \nu_i}\overline{\boldsymbol{P}}^k g_k(s,t) + \overline{\boldsymbol{P}}^k \frac{\partial}{\partial \nu_i} g_k(s,t)\right].$$

Now the parameter ν_i will either belong to the constant matrix $\boldsymbol{Q}$ or it will belong to the time-variant scalar function $f(t)$. The previous result may therefore be further simplified for either one of both cases:

$$\frac{\partial}{\partial \nu_i}\boldsymbol{P}(s,t) = \begin{cases} \displaystyle\sum_{k=0}^{\infty} \frac{\partial}{\partial \nu_i}\overline{\boldsymbol{P}}^k g_k(s,t), & \text{if } g_k \text{ is not a function of } \nu_i, \text{ or} \\ \displaystyle\sum_{k=0}^{\infty} \overline{\boldsymbol{P}}^k \frac{\partial}{\partial \nu_i} g_k(s,t), & \text{if } \overline{\boldsymbol{P}} \text{ is not a function of } \nu_i. \end{cases}$$

References

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5:141–150.

AASHTO. (2005). *Pontis Release 4.4 Technical Manual*. AASHTO, Washington, D.C..

Abaza, K. A., Ashur, S. A. and Al-Khatib, I. A. (2004). Integrated pavement management system with a Markovian prediction model. *Journal of Transportation Engineering*, 130(1):24–33.

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover Publications, New York, NY.

Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about Markov chains. *Annals of Mathematical Statistics*, 28:89–110.

Baik, H.-S., Jeong, H. S. and Abraham, D. M. (2006). Estimating transition probabilities in Markov chain-based deterioration models for management of wastewater systems. *Journal of Water Resources Planning and Management*, 132(1):15–24.

Bickenbach, F. and Bode, E. (2003). Evaluating the Markov property in studies of economic convergence. *International Regional Science Review*, 26(3):363–392.

Billingsley, P. (1961). Statistical models in markov chains. *Annals of Mathematical Statistics*, 32:12–40.

Black, M., Brint, A. T. and Brailsford, J. R. (2005a). Comparing probabilistic methods for the asset management of distributed items. *Journal of Infrastructure Systems*, 11(2):102–109.

Black, M., Brint, A. T. and Brailsford, J. R. (2005b). A semi-Markov approach for modelling asset deterioration. *Journal of the Operational Research Society*, 56:1241–1249.

Bulusu, S. and Sinha, K. C. (1997). Comparison of methodologies to predict bridge deterioration. *Transportation Research Record*, 1597:34–42.

Butt, A. A., Shahin, M. Y., Feighan, K. J. and Carpenter, S. H. (1987). Pavement performance prediction model using the Markov process. *Transportation Research Record*, 1123:12–19.

Cameron, A. C. and Travedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press, Cambridge, United Kingdom.

Cappé, O., Buchoux, V. and Moulines, E. (1998). Quasi-newton method for maximum likelihood estimation of hidden markov models. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), May 12-15, 1998, Seattle WA*, pages 2265–2268. IEEE.

Carnahan, J. V., Davis, W. J., Shahin, M. Y., Keane, P. L. and Wu, M. I. (1987). Optimal maintenance decisions for pavement management. *Journal of Transportation Engineering*, 113(5):554–572.

Cesare, M., Santamarina, J. C., Turkstra, C. J. and Vanmarcke, E. (1994). Risk-based bridge management: optimization and inspection scheduling. *Canadian Journal of Civil Engineering*, 21(6):897–902.

Châtelet, E., Djebabra, M., Dutuit, Y. and dos Santos, J. (1996). A comparison between two indirect exponentiation methods and O.D.E. integration method for RAMS calculations based on homogeneous Markovian models. *Reliability Engineering and System Safety*, 51(1):1–6.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis*, 5:315–327.

Corotis, R. B., Ellis, J. H. and Jiang, M. (2005). Modeling of risk-based inspection, maintenance and life-cycle cost with partially observable Markov decision processes. *Structure & Infrastructure Engineering*, 1(1):75–84.

Dacunha, J. J. (2005). Transition matrix and generalized matrix exponential via the Peano-Baker series. *Journal of Difference Equations and Applications*, 11(15):1245–1264.

Dekker, R. (1996). Applications of maintenance optimization models: a review and analysis. *Reliability Engineering and System Safety*, 51(3):229–240.

DeStefano, P. D. and Grivas, D. A. (1998). Method for estimating transition probability in bridge deterioration models. *Journal of Infrastructure Systems*, 4(2):56–62.

Ellis, J. H., Jiang, M. and Corotis, R. B. (1995). Inspection, maintenance, and repair with partial observability. *Journal of Infrastructure Systems*, 1(2):92–99.

Faddy, M. J. (1995). Phase-type distributions for failure times. *Mathematical and computer modelling*, 22(10–12):63–70.

FHWA. (1995). *Recording and coding guide for the structure inventory and appraisal of the nation's bridges. Report No. FHWA-PD-96-001.* U.S. Department of Transportation, Federal Highway Administration, Washington, D.C..

Fisher, R. A. (1912). On a absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A*, 222:309–368.

Flehinger, B. J. (1962). A Markovian model for the analysis of the effects of marginal testing on system reliability. *Annals of Mathematical Statistics*, 33(2):754–766.

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

Gaal, G. C. M. (2004). *Prediction of deterioration of concrete bridges: corrosion of reinforcement due to chloride ingress and carbonation*. PhD thesis, Delft University of Technology, Delft, Netherlands.

Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18(4):1501–1555.

Golabi, K. (1983). A markov decision modeling approach to a multi-objective maintenance problem. In *Essays and Surveys on Multiple Criteria Decision Making*, pages 115–125, New York, NY. Springer.

Golabi, K., Kulkarni, R. B. and Way, G. B. (1982). A statewide pavement management system. *Interfaces*, 12(6):5–21.

Golabi, K. and Pereira, P. (2003). Innovative pavement management and planning system for road network of portugal. *Journal of Infrastructure Systems*, 9(2):75–80.

Golabi, K. and Shepard, R. (1997). Pontis: a system for maintenance optimization and improvement of US bridge networks. *Interfaces*, 27(1):71–88.

Heidelberger, P. and Goyal, A. (1988). Sensitivity analysis of continuous time Markov chains using uniformization. In Iazeolla, G., Courtois, P. J. and Boxma, O. J., editors, *Proceedings of the Second International MCPR Workshop, Rome, Italy, May 25-29, 1987*, pages 93–104, Amsterdam. Elsevier Science (North-Holland).

Howard, R. A. (1971). *Dynamic Probabilistic Systems, Volume II: Semi-Markov and Decision Processes*. John Wiley & Sons, New York.

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W. and Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209.

Jaynes, E. T. (1957). Information theory and statistical mechanics I. *Physical Review*, 106:620–630.

Jennrich, R. I. and Bright, P. B. (1976). Fitting systems of linear differential equations using computer generated exact derivatives. *Technometrics*, 18(4):385–392.

Jia, X. and Christer, A. H. (2002). A prototype cost model of functional check decisons in reliability-centred maintenance. *Journal of the Operational Research Society*, 53(12):1380–1384.

Jiang, M., Corotis, R. B. and Ellis, J. H. (2000). Optimal life-cycle costing with partial observability. *Journal of Infrastructure Systems*, 6(2):56–66.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous univariate distributions, Volume 1*. John Wiley & Sons, 2nd edition.

Kailath, T. (1980). *Linear systems*. Prentice-Hall, Englewood Cliffs.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871.

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

Kallen, M. J. and van Noortwijk, J. M. (2005b). A study towards the application of Markovian deterioration processes for bridge maintenance modelling in the Netherlands. In Kołowrocki, K., editor, *Advances in Safety and Reliability: Proceedings of the European Safety and Reliability Conference (ESREL 2005), Tri City (Gdynia-Sopot-Gdánsk), Poland, 27-30 June, 2005*, pages 1021–1028, Leiden, Netherlands. Balkema.

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

Kallen, M. J. and van Noortwijk, J. M. (2006b). Statistical inference for Markov deterioration models of bridge conditions in the Netherlands. In Cruz, P. J. S., Frangopol, D. M. and Neves, L. C., editors, *Bridge Maintenance, Safety, Management, Life-Cycle Performance and Cost: Proceedings of the Third International Conference on Bridge Maintenance, Safety and Management, 16-19 July, Porto, Portugal, 2006.*, pages 535–536, London. Taylor & Francis.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42:855–865.

Kleiner, Y. (2001). Scheduling inspection and renewal of large infrastructure assets. *Journal of Infrastructure Systems*, 7(4):136–143.

Kohlas, J. (1982). *Stochastic methods of operations research*. Cambridge University Press, Cambridge, United Kingdom.

Lee, T. C., Judge, G. G. and Zellner, A. (1970). *Estimating the parameters of the Markov probability model from aggregate time series data*. North-Holland, Amsterdam.

Lindsey, J. K. (1996). *Parametric Statistical Inference*. Oxford University Press, Oxford, United Kingdom.

Long, S. J. (1997). *Regression models for categorical and limited dependent variables*. Sage Publications, Thousand Oaks, CA.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall, London.

Madanat, S., Mishalani, R. and Wan Ibrahim, W. H. (1995). Estimation of infrastructure transition probabilities from condition rating data. *Journal of Infrastructure Systems*, 1(2):120–125.

Madanat, S. and Wan Ibrahim, W. H. (1995). Poisson regression models of infrastructure transition probabilities. *Journal of Transportation Engineering*, 121(3):267–272.

Madanat, S. M. (1993). Optimal infrastructure management decisions under uncertainty. *Transportation Research: Part C*, 1(1):77–88.

Madanat, S. M., Karlaftis, M. G. and McCarthy, P. S. (1997). Probabilistic infrastructure deterioration models with panel data. *Journal of Infrastructure Systems*, 3(1):4–9.

Micevski, T., Kuczera, G. and Coombes, P. (2002). Markov model for storm water pipe deterioration. *Journal of Infrastructure Systems*, 8(2):49–56.

Mishalani, R. G. and Madanat, S. M. (2002). Computation of infrastructure transition probabilities using stochastic duration models. *Journal of Infrastructure Systems*, 8(4):139–148.

Moler, C. and van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20(4):801–836.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, Singapore, 3rd edition.

Morcous, G. (2006). Performance prediction of bridge deck systems using Markov chains. *Journal of the Performance of Constructed Facilities*, 20(2):146–155.

Moubray, J. (1997). *Reliability-Centred Maintenance*. Industrial Press, New York, 2nd edition.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. John Hopkins University Press, Baltimore, MD.

Phares, B. M., Washer, G. A., Rolander, D. D., Graybeal, B. A. and Moore, M. (2002). Routine highway bridge inspection condition documentation accuracy and reliability. *Journal of Bridge Engineering*, 9(4):403–413.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rindos, A., Woolet, S., Viniotis, J. and Trivedi, K. (1995). Exact methods for the transient analysis of nonhomogeneous continuous time Markov chains. In Stewart, W. J., editor, *2nd International Workshop on the Numerical Solution of Markov Chains*, pages 121–133, Boston, MA. Kluwer Academic Publishers.

Roelfstra, G., Hajdin, R., Adey, B. and Brühwiler, E. (2004). Condition evolution in bridge management systems and corrosion-induced deterioration. *Journal of Bridge Engineering*, 9(3):268–277.

Ross, S. M. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco, CA.

Ross, S. M. (2000). *Introduction to Probability Models*. Harcourt/Academic Press, Burlington, MA, 7th edition.

Scherer, W. T. and Glagola, D. M. (1994). Markovian models for bridge maintenance management. *Journal of Transportation Engineering*, 120(1):37–51.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Skuriat-Olechnowska, M. (2005). Statistical inference and hypothesis testing for Markov chains with interval censoring: application to bridge condition data in the Netherlands. Master's thesis, Delft University of Technology, Delft, Netherlands.

Smilowitz, K. and Madanat, S. M. (2000). Optimal inspection and maintenance policies for infrastructure systems under measurement and prediction uncertainty. In *Transportation Research Circular 489: Presentations from the 8th International Bridge Management Conference, April 26-28, 1999, Denver, Colorado*. Transportation Research Board.

Sztul, M. (2006). Modelling uncertainty in inspections of highway bridges in the Netherlands. Master's thesis, Delft University of Technology, Delft, Netherlands.

Tijms, H. C. (2003). *A first course in stochastic models*. Wiley, Chichester, United Kingdom.

Tsai, H. and Chan, K. S. (2003). A note on parameter differentiation of matrix exponentials, with applications to continuous-time modelling. *Bernouilli*, 9(5):895–919.

van Moorsel, A. P. A. and Wolter, K. (1998). Numerical solution of non-homogeneous markov processes through uniformization. In Zobel, R. N. and Möller, D. P. F., editors, *12th European Simulation Multiconference - Simulation - Past, Present and Future, June 16-19, 1998, Machester, United Kingdom*, pages 710-717. SCS Europe.

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

van Noortwijk, J. M. (2007). A survey of the application of gamma processes in maintenance. *Reliability Engineering and System Safety*, in press. doi:10.1016/j.ress.2007.03.019.

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

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

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

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *The Annals of Statistics*, 28(3):779–814.

Wilcox, R. M. (1967). Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics*, 8(4):962–982.

Wirahadikusumah, R., Abraham, D. and Iseley, T. (2001). Challenging issues in modeling deterioration of combined sewers. *Journal of Infrastructure Systems*, 7(2):77–84.

Woodward, R. J. and others (2001). BRIME - bridge management in Europe. Technical Report, Transport Research Laboratory (TRL), United Kingdom. Project funded by the European Commission under the RTD programme of the 4th framework programme.

Yang, J., Gunaratne, M., Lu, J. J. and Dietrich, B. (2005). Use of recurrent Markov chains for modeling the crack performance of flexible pavements. *Journal of Transportation Engineering*, 131(11):861–872.

Acknowledgments

At the end of a four year endeavour, there are many people to thank for their support and cooperation during the research for this thesis. First, I would like to thank the two people who have been most instrumental to this project: Jan van Noortwijk and Leo Klatter. In many aspects, this research would not have been possible without them. I'm very grateful for their commitment to the project and for the countless discussions which have ultimately shaped the contents of this thesis. In their role as members of the supervisory committee, they did not shun a healthy dosis of criticism. Neither did the other members, being Jaap Bakker, Jan Stijnen and Hans van der Weide, and I thank each of them for their constructive comments and their research suggestions. The best way to take a broader view on things, is by having other people comment on your work. Finally, during the final stages of preparing this thesis, several members of the Ph.D. committee have made valuable suggestions towards improving the manuscript for which I'm very grateful.

I would like to thank all my colleagues at HKV for a very pleasant working environment. I was fortunate to meet many interesting people at the Bouwdienst in Utrecht and my days working there have been a very nice experience. Many thanks to everyone on the 6th floor of the faculty of EEMCS in Delft. It has been lots of fun, at times quite hilarious, and I'm sure I will always remember my time in Delft as one of the most enjoyable and challenging periods of my life. There are too many people to mention here, but I would particularly like to thank those people who have supported me with some of the practical issues during my research: Geke in Utrecht, Cindy and Carl in Delft, and Ellen, Melanie, Carina, Charissa and Gina in Lelystad. Finally I would also like to thank my paranymphs Daniel and Sebastian.

Over the past four years, I've had the pleasure to assist a number of students during their research for their Master's thesis. I thank Monika Skuriat-Olechnowska, Magda Sztul and Jojanneke Dirksen for their great work and pleasant cooperation.

Finally, I thank the people who are dearest to me. My parents for their love and for helping me get to where I am now. My wife Antje and my daugther Lena ... life wouldn't even nearly be as enjoyable without you. Your love and moral support have made this project all the more special.

Maarten-Jan Kallen
The Hague, October 13, 2007

About the author

Maarten-Jan Kallen was born in Creve Coeur, a community in St. Louis County in the state of Missouri (U.S.A.), on February 16th, 1978.

In 1996, he obtained his diploma of Secondary Education at the Lutgardiscollege in Oudergem, a community in Brussels, Belgium. The fall 1996 and spring 1997 semesters he was a freshman at the University of Colorado in Boulder, with a major in Applied Mathematics and a minor in Computer Science. Starting in the fall semester of 1997, he studied Applied Mathematics at the Delft University of Technology in Delft, the Netherlands. He obtained his M.Sc. diploma in June, 2003, with the thesis entitled 'Risk Based Inspection in the Process and Refining Industry'. It was completed as part of an internship at Det Norske Veritas BV in Rotterdam, the Netherlands.

Between April 2003 and April 2007, Maarten-Jan was employed as a researcher in the field of maintenance optimization at HKV Lijn in Water BV in Lelystad, the Netherlands. The research performed during this period was financially supported by the Civil Engineering Division (special services division of the Directorate-General for Public Works and Water Management) and the faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology. As of April 2007, he is a consultant in the Risk and Safety group at HKV Lijn in Water BV.

Maarten-Jan is married to Antje Marcantonio. They have a daughter Lena Marie.