MAINTENANCE MEETS PRODUCTION

On the Ups and Downs of a Repairable System

Promotor:

prof. dr. A. van Harten

Overige leden van de promotiecommissie:

prof. dr. W. Albers
prof. drs. P. Bouw
prof. dr. R. Dekker
prof. dr. A.G. de Kok
dr. ir. L. Pintelon
prof. dr. W.H.M. Zijm

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Maintenance meets production: on the ups and downs of a repairable system / Gerhard Christiaan van Dijkhuizen. - [S.l. : s.n.], 1998 (Enschede : Print Partners Ipskamp). - 184 p. : fig., tab. ; 24 cm. - Proefschrift Universiteit Twente, Enschede. - Met samenvatting in het Nederlands. - Met lit. opg.

ISBN 90-365-1178-X

MAINTENANCE MEETS PRODUCTION

On the Ups and Downs of a Repairable System

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 23 oktober 1998 te 15.00 uur.

door

Gerhard Christiaan van Dijkhuizen
geboren op 22 september 1970
te Amstelveen

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. A. van Harten

“In the beginner’s mind
there are many possibilities,
but in the expert’s
there are few.”

Shunryu Suzuki Roshi (1905-1971)

Voorwoord

Dit proefschrift bevat de meest tastbare resultaten van ruim vier jaar promotie-onderzoek, dat is verricht bij de vakgroep Operationele Methoden en Systeemtheorie (OMST) van de faculteit Technologie & Management (T&M) aan de Universiteit Twente (UT). Zoals de titel van het proefschrift reeds doet vermoeden, heeft mijn onderzoek zich toegespitst op de mogelijke interacties tussen onderhoud en produktie (meet = ontmoeten), en heb ik daarbinnen vooral de rol van de onderhoudsfunctie belicht (meet = tegemoet komen). Met behulp van een tweetal veelgebruikte termen uit het vakgebied (up = hij doet het, down = hij doet het niet), geeft de subtitel vervolgens aan dat de verstandhouding tussen onderhoud en produktie in de praktijk nog regelmatig te wensen over laat.

Vele mensen hebben op één of andere wijze een bijdrage geleverd aan de totstandkoming van dit proefschrift. Ik wil *met name* mijn promotor Aart van Harten bedanken voor de vele uren die hij aan het kritisch lezen van mijn hersenspinsels heeft besteed, en de vaak nuttige suggesties die daaruit voortvloeiden. Ook aan alle overige collega's van de vakgroep OMST, de faculteit T&M en mijn tijdelijke werkgever KLM ben ik dank verschuldigd. Hun aanwezigheid heeft mij de afgelopen jaren doorgaans met plezier naar mijn werk doen gaan. Evenzeer wil ik familie, vrienden en kennissen bedanken voor de soms zichtbare, maar veelal onzichtbare steun die zij hebben geleverd. Tenslotte is een woord van dank aan spelers en begeleiders van de voetbalverenigingen Drienerlo en Sparta op zijn plaats, omdat ze niet zelden de ideale uitlaatklep bleken te zijn voor tijdens het werk opgedane frustaties.

Terug- en vooruitblikkend besef ik als geen ander dat er maar weinigen zijn die dit proefschrift van kop tot staart zullen danwel kunnen doorspitten. Dat wil echter nog niet zeggen dat al mijn inspanningen voor niets zijn geweest. Binnen afzienbare tijd zal vrijwel het gehele proefschrift in één of andere vorm als wetenschappelijke publicatie in de vakliteratuur zijn verschenen, en zijn de resultaten ervan in principe dus voor iedereen toegangelijk. Nog belangrijker echter is dat ik mijzelf in de afgelopen jaren regelmatig ben tegengekomen, en als gevolg daarvan over meer zelfkennis beschik dan ooit tevoren. Spookverhalen als zou promoveren je vier jaar

van je leven kosten, en je geen steek verder brengen in het bedrijfsleven, zijn wat mij betreft dan ook volkomen uit de lucht gegrepen.

Het zal menigeen dan ook nauwelijks verbazen dat ik de academische wereld voorlopig vaarwel zal zeggen. Hoewel ik met een tevreden gevoel terugkijk op een vruchtbare combinatie van ruim viereneenhalf jaar onderwijs en onderzoek, ben ik ervan overtuigd over een aantal jaren spijt te zullen krijgen als ik niet nu de stap naar het bedrijfsleven maak. Een kijkje in de keuken van zowel de toegepaste wiskunde als de technische bedrijfskunde, hebben mij tot de conclusie doen komen dat juist op de raakvlakken van deze vakgebieden nog veel interessant werk te doen is. Desalniettemin spreek ik de hoop uit nog voldoende tijd te kunnen vrij maken om zo nu en dan, en bij voorkeur met wat oude collega's, een artikeltje in elkaar te knutselen.

Gerhard van Dijkhuizen Enschede, oktober 1998

Contents

Chapter 1

Introduction

The British Standards Institute defines maintenance as the combination of all technical and associated administrative actions intended to retain an item in, or restore it to, a state in which it can perform its required function. In production systems, the objective of maintenance is to ensure that the underlying equipment performs regularly and efficiently, by reducing the possibilities of breakdowns or failures, and by minimizing the production loss resulting from them. Simply stated, maintenance management attempts to maximize the performance of a production system, while keeping overall maintenance efforts in terms of the associated time and/or costs at an acceptable level. Although this statement sounds relatively simple, maintenance management addresses a rich and complex problem area. Therefore, let us start with an impression of the sources of complexity associated with maintenance management, and discuss the contributions of our work to existing literature.

1.1 Scope of this thesis

Unfortunately, almost all production systems in our society are subject to random failure of one or more of their components. In general, these failures may severely affect the performance and profitability of such systems, since they often occur at inconvenient times, and usually involve high recovery and consequential costs. Nowadays, it is widely recognized that the influence of failures on system performance is not just to be taken for granted. More than ever, manufacturing industries have realized that there is a hugh potential of efficiency improvements, if the number of component failures could be reduced and/or components could be repaired or replaced before they fail. In the past few decades, this has given rise to an increased effort to introduce technological innovations (e.g. design for reliability, design for maintainability),

and better preventive maintenance concepts in practice as well. At the same time, a growing interest can be observed in theoretical literature concerning the modelling and optimization of maintenance and reliability in failure prone systems.

Traditionally, production and operations managers have rarely viewed the maintenance function as a competitive factor in their firm's business strategies. Nowadays, as many industries are moving towards just-in-time production, and at the same time rely on highly mechanized and automated production systems and processes, the strategic importance of maintenance is widely recognized. More than ever, maintenance is being considered as a basis factor to satisfy production needs, rather than a necessary evil. Therefore, it should be managed together with production on an equal basis, with an open eye for their interactions. Nevertheless, maintenance interventions are still too often considered as constraints on production planning and scheduling, or the other way around. Ideally, however, maintenance jobs should be treated as capacity-consuming production jobs, to be scheduled on resources of limited capacities. In this respect, there is a perspective of significant gains if maintenance and production activities are considered simultaneously, not only in an operational planning phase, but also at a tactical and strategical level.

Strategic maintenance planning is concerned with decisions that are aimed at keeping a company succesful on a long term basis. Typical examples are decisions whether or not to replace or upgrade certain production equipment, whether or not to outsource several maintenance activities, etcetera. Tactical maintenance planning is concerned with medium term decision making, and its primary objective is to ensure the effective and efficient use of production equipment and/or spare parts to assure a specified performance level. Finally, operational maintenance planning is concerned with priority setting, coordination and execution of preventive and/or corrective maintenance activities, as well as possible interactions with production scheduling. Complicating factors in this respect are the limited time that is usually available for maintenance activities in view of production needs, and the uncertainties with respect to the occurence of, and the time required for these activities.

This thesis is another contribution to the further development of mathematical models for maintenance optimization at a tactical level, which systematically and explicitly take into account interactions with production in several dimensions. More specifically, our overall research objective can be stated as follows:

> to develop **mathematical models** which can **assist** in the **optimization** of **maintenance policies** for **complex systems**, thereby taking into account **interactions with production** in terms of **technical process reliability**, **system availability**, and **minimization of costs**.

Basically, the objective of mathematical models for maintenance management is to provide a quantitative assessment of both maintenance costs and benefits, in order to arrive at an optimal balance between the two. In this respect, a clear distinction must be made between preventive maintenance (before failure), predictive maintenance (just before failure), and corrective maintenance (upon or after failure). In general, predictive maintenance strategies rely on a deeper understanding of failure causes,and require a so-called failure indicator as well (e.g. crack growth, vibration analysis, thermography). In this thesis, we will restrict ourselves to preventive and corrective maintenance strategies.

Usually, a preventive maintenance action requires less time and costs compared with an analagous corrective maintenance action. Moreover, preventive maintenance actions can often be planned in advance, whereas corrective maintenance actions usually occur at inconvenient times. Apparently, there is a potential of both cost and time reductions by conducting maintenance preventively rather than correctively. On the other hand, too frequent preventive maintenance can be inefficient, or even ineffective too. Simply stated, the models presented in this thesis aim at providing maintenance management decision support in this respect. Let us now consider in some more depth the scope of the problem areas we are going to address.

As a starting point, the complexity of a production system does not only relate to its (potential) relations and interactions with other systems (e.g. intermediate buffers, safety stocks, standby equipment), but also to the maintainability of the system itself. Amongst several other factors, the maintainability of a production system can be expressed in terms of the preparatory set-up activities that have to be carried out, before actual maintenance actions can take place. As a consequence, there is a perspective of significant savings if maintenance activities are carried out simultaneously. Traditionally, maintenance optimization models have accounted for these economies of scale, by assuming that a fixed set-up cost is incurred at each occasion for preventive and/or corrective maintenance. In this thesis, we will present a new, much richer and more powerful modelling framework, which allows for the coordination of preventive maintenance activities in a multi-component production system with **multiple interrelated set-up activities**.

The technical process reliability relates to the extent in which a production system is able to perform its required function. Obviously, this ability is strongly influenced by the effectiveness of the underlying maintenance concept. In this respect, it is not uncommon for preventive maintenance frequencies to be determined in advance, e.g. by specialized maintenance engineers, or due to safety restrictions and/or legislation. This is particularly true for highly regulated production environments, such as air-

lines, nuclear power plants, and offshore platforms. In such cases, these superimposed maintenance frequencies are usually, or must be treated as constraints in further optimization techniques. In this thesis, we will focus on **frequency-constrained** maintenance jobs with fixed costs, as well as maintenance jobs with more general **frequency-dependent costs**.

Although maintenance costs are usually expressed in terms of a long run average, there is a strong difference between the long and short term behavior of a production system, in view of the down times due to maintenance. In literature, this difference has been recognized and incorporated by making a clear distinction between the **limiting availability**, and the **interval availability distribution** of a production system. Simply stated, the limiting availability reflects the average performance of a production system over an infinite period of time, whereas the interval availability distribution refers to its actual performance during a finite time interval. In general, it depends on the type of production environment whether the limiting availability or the interval availability is the most appropriate performance measure to be used. Nevertheless, the impact of preventive maintenance on both performance measures is obvious, but the latter is not so well explored in existing literature. It will be studied thoroughly further on in this thesis.

Another important aspect concerning the availability of a production system, is whether and to which extent the down times associated with preventive and/or corrective maintenance involve production losses. If this effect is strong, as is the case with so-called bottleneck machines, the consequential costs of maintenance are usually significantly larger. In general, corrective maintenance requires an interruption of the production process, whereas preventive maintenance can be planned at more convenient times (e.g. between shifts, at night, or during weekends), or at so-called **maintenance opportunities** (e.g. idle times, withdrawn orders, machine failures). Since these opportunities are usually not known in advance, or at best on a short term basis, there is a potential of both time and cost reductions if some flexibility is build in concerning the starting time of preventive maintenance. Although this is a widespread common sense in practice, it certainly is an underexposed point of view in existing literature. In this thesis, we will incorporate this kind of flexibility into some elementary maintenance models.

The other way around, it is also possible that maintenance activities, either preventive or corrective, must be carried out during predefined time intervals, or so-called **maintenance slots**. Typical examples of this type can be found in airline companies, where a given number of flights must be realized with a given number of aircrafts, and all maintenance activities must be carried out in between these flights.

In such cases, it must be decided how many maintenance slots of which type must be available within the timetable, and how many maintenance engineers of which type must be assigned to these slots, in order to facilitate maintenance complying with the constraints set by higher management. Complicating factors in this respect are the uncertainty associated with the occurence of corrective maintenance jobs, as well as the variation in corresponding repair times and due dates. In this thesis, we will present the results of a case study that was carried out at the Line Maintenance department of KLM Royal Dutch Airlines at Schiphol Airport.

Summarizing, the main contribution of the models presented in this thesis, is that they exploit the advantages of preventive maintenance in view of production needs, more systematically and explicitly than in existing literature, at least up to our knowledge. More specifically, the main problem areas addressed in this thesis can be summarized as follows:

- the coordination of preventive maintenance activities in a multi-component production system with multiple interrelated set–up activities,
- the influence of preventive maintenance strategies on the interval availability distribution of an unreliable production system,
- the potential savings of building in some flexibility concerning the actual starting time of preventive maintenance in an operational planning phase,
- the impact of maintenance slots in a practical context of corrective maintenance activities with several sources of uncertainty and variation.

These problem areas, and the associated research questions, will be addressed more thoroughly in chapters 2 to 7 respectively. In the remainder of this first chapter, we will focus on clarification and specification of the various notions used in this thesis, and a further positioning of our work within the literature. As a starting point, section 1.2 contains a more detailed discussion of the interactions between maintenance and production, and identifies some interesting research opportunities as well. Subsequently, section 1.3 provides an elementary introduction into the most fundamental aspects of maintenance management, and puts the role of maintenance optimization models in a wider perspective. In section 1.4, a brief review on mathematical models for maintenance optimization is presented, and our work is related to existing literature. Finally, section 1.5 gives an overview of this thesis.

1.2 Maintenance meets production

It was not so long ago that price and quality were the only competitive factors, and customer satisfaction could be ensured by maintaining large inventories of finished products. Nowadays, rapid technological changes and smaller profit margins have made such a strategy uneconomical, literally forcing companies to run with lower inventory levels. Moreover, customers have changed in the sense that they expect high quality products, quick response to orders, fast and reliable product delivery and seamless service, all against reasonable costs. Under this increasing pressure, manufacturing firms are being forced to improve continuously in several dimensions, of which the most fundamental are unit cost efficiency, product quality, and timeliness of delivery (Hopp and Spearman 1996).

Although these dimensions are broadly applicable to a lot of manufacturing industries, their relative importance varies from one firm to another. For example, a manufacturer of a commodity (e.g. sugar, coffee) depends critically on efficiency, since low cost is a condition for survival. On the other hand, a manufacturer of premium goods (e.g. automobiles, watches) mainly relies on quality to retain its market. Finally, a manufacturer of high-tech products (e.g. computers, televisions) requires speed of introduction in order to be competitive, and to maximally exploit potential profits within the limited economic lifecycle of its products. Clearly, maintenance management plays an important role in each of these industries, and in each of the above-mentioned dimensions.

In this respect, it is not only the reliability and availability of a production system, but also its variability that matters (see Figure 1.1). In literature, the **reliability** of a production system is often expressed in terms of the probability that the system will operate satisfactorily (i.e. without failures) for at least a certain period of time. In a similar way, the **availability** of a production system is usually defined as the probability that the system will be operating satisfactorily at an arbitrary point in time, or equivalently as the long run average fraction of time that the system is operational. Finally, the **variability** or predictability of a production system reflects its ability to produce at a more or less constant rate, and it is often expressed in terms of an interval availability distribution. In this thesis, we will mainly be concerned with the availability (long term behavior) and variability (short term behavior) of unreliable production systems. But first, let us discuss in some more detail the potential benefits of maintenance management in each of the above-mentioned dimensions: unit cost efficiency, product quality, and timeliness of delivery.

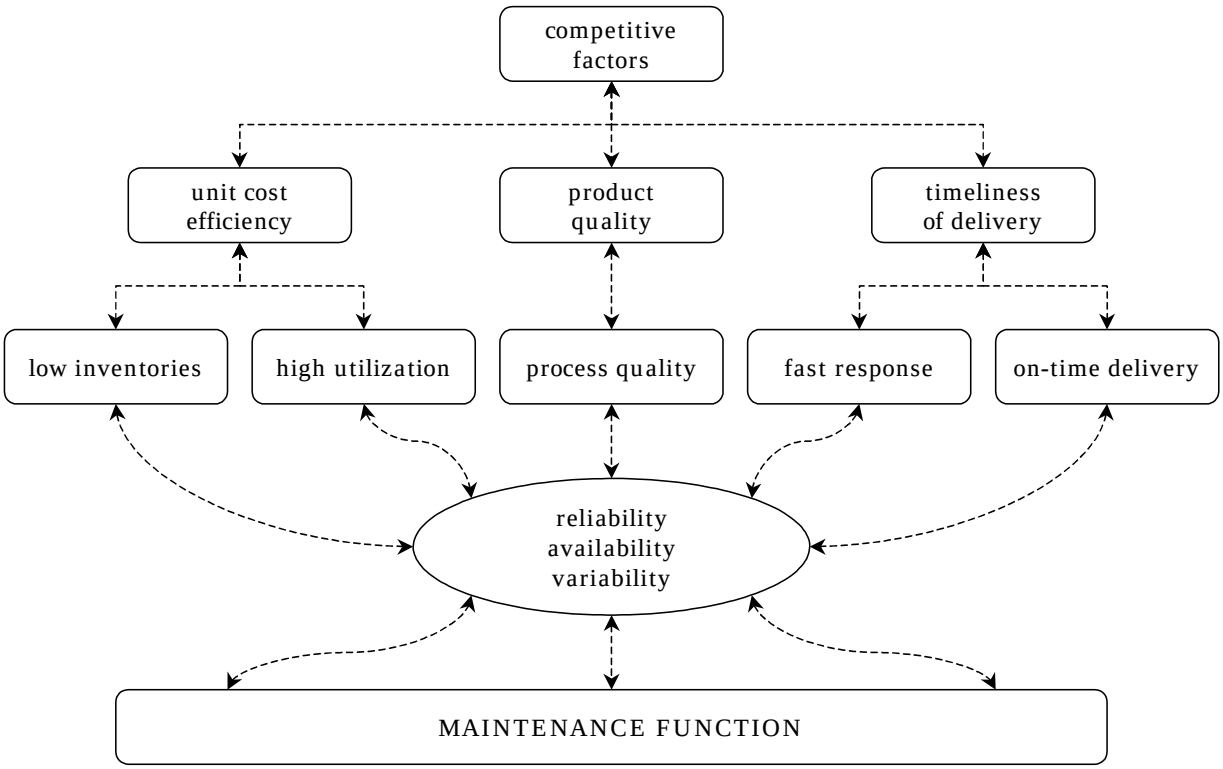


Figure 1.1: Maintenance as a contributive factor to logistic performance.

1.2.1 Unit cost efficiency

Efficient utilization of the available resources (e.g. labour, material, equipment) has always been essential in view of keeping operating costs at a competitive level. From the customer standpoint, it is unit cost (total costs divided by total volume) that matters, implying that bost cost reduction and volume enhancement are commercially worthy objectives. In this respect, it is not only the frequency, but also the timing of maintenance that matters. In general, preventive maintenance involves lower cost and less time compared with corrective maintenance. Moreover, preventive maintenance can be carried out at more convenient times (e.g. at night, in weekends, during holidays), or at so-called maintenance opportunities (e.g. idle times, machine failures, withdrawn orders), whereas corrective maintenance usually requires interruption of the production process. Simply stated, these advantages of preventive maintenance provide the rationale for the models presented in this thesis.

By replacing or revising components before they fail, production systems might be prevented from suddenly breaking down, thereby avoiding high corrective maintenance costs, and long and expensive down times. This effect is even stronger if mutual dependencies between components are taken into account. In some cases, the

failure of a component may increase the failure rate of other components (functional dependence), or require the replacement of non-failed components as well (structural dependence). Moreover, preventive maintenance on different components can often be grouped into maintenance packages to reduce set-up times and/or costs (economic dependence), whereas this is somewhat more complicated in case of corrective maintenance. In this thesis, we will restrict ourselves to economic dependencies between components.

In view of efficiency, another advantage of preventive maintenance is that it allows for the reduction or elimination of intermediate buffers and safety stocks, that are usually maintained to keep production going if one or more machines have failed. In a similar way, the number of spare parts held in stock can often be reduced significantly, since the majority of repairs and replacements can be planned in advance. Apparently, there is a perspective of significant savings in inventory holding costs, if maintenance is carried out preventively rather than correctively. In this thesis, such considerations will be left out of consideration. Nevertheless, they could be incorporated implicitly in some of our models.

1.2.2 Product quality

The past few decades have brought widespread recognition that quality is also a key competitive weapon. Although external or product quality has always been a concern in manufacturing industries, the quality revolution of the 1980's served to focus attention on internal or process quality at each step in the production process, and its relationship to customer satisfaction. Facets of operations management, such as statistical process control, have loomed largely in this context as components of Total Quality Management (Hakes 1991). Since the degree to which a product conforms to its technical specifications is strongly related to the capabilities of the underlying production equipment, the contribution of maintenance to both product and process quality is nowadays widely recognized. These interactions, however, will not be explicitly addressed in this thesis.

1.2.3 Timeliness of delivery

While cost and quality remained cricitical as always, the 1990's have become the decade of speed. Rapid development of new products, coupled with fast and on-time delivery, are the pillars of manufacturing strategies adopted in many different industries. Responsive delivery, without inefficient excess inventory, requires short cycle times, reliable processes, and effective integration of disparate functions (e.g.

maintenance and production). In this respect, there is a perspective of significant improvements in system efficiency, if the variabilities in the production processes can be reduced. In many plants, unscheduled downtimes due to random breakdowns are one of the largest, and most disruptive sources of variability.

Ideally, a production system should be able to produce at a more or less constant rate (i.e. without service interruptions), while retaining a satisfactory production capacity in the long run. Therefore, a production system with frequent, predictable and short interruptions is to be preferred above one with infrequent, unpredictable and long interruptions, all other things being equal. In many practical situations, and in most capacity planning systems, it is not only the average production capacity in the long run (limiting availability), but also the guaranteed production capacity during a finite period of time (interval availability) that matters. This is a potentially valuable insight, since in practice the variability of a production system may be reduced by conducting preventive maintenance at regular intervals. Therefore, it will be studied thoroughly in this thesis.

1.3 Maintenance management

In the past few decades, production and operations managers have been confronted with a variety of innovative and revolutionary concepts, among which some of the most famous are MRP (Manufacturing Resource Planning), JIT (Just In Time), and OPT (Optimized Production Technology). Although each of them has undoubtedly provided useful insights, nowadays manufacturing systems are facing problems which are far too complicated to be tackled by buzzword management. Effective managers of the future will have to rely on a solid understanding of their systems, in order to identify opportunities for improvement (Hopp and Spearman, 1996). In the current competitive environment of short lead times and on-time deliveries, this means that maintenance management plays a key strategic role in the optimization of business processes.

1.3.1 Historical perspective

It was not so long ago, however, that maintenance was simply regarded as an unavoidable and unpredictable part of production. There was no department responsible for the maintenance function, and the majority of maintenance activities was of a corrective nature. As a consequence, there was a lot of uncertainty with respect to the production process due to random breakdowns, causing completion times and

product quality to be highly unreliable. In this respect, a revolutionary change in the maintenance function was bound to happen, and maintenance finally became the responsibility of a special department. About three degades ago, companies began to realize that engineering qualities alone were no longer sufficient for supervising the maintenance department, and maintenance management was born. Today, maintenance management is far more important than it has ever been before.

Historically, the typical size of the maintenance department in a manufacturing organization ranged from 5 to 10 percent of the operating force. Today, the proportional size of the maintenance effort compared to the production effort (including outsourcing) has increased, and is projected to increase even further. This tendency is caused by the ongoing transformation of labour-intensive into capital-intensive plants, viz. the mechanization and automation of production processes. This trend has decreased the need for operators, and at the same time resulted in a greater demand for technicians, electricians, and other service people. In e.g. refineries, it is not uncommon that the maintenance department represents about 30 percent of total manpower, and overall maintenance costs are the largest part of the operational budget.

Nowadays, it is often found cost-effective to leave the day-to-day or routine maintenance activities (e.g. lubrication, cleaning, monitoring) in the hands of machine operators, since they often know best when their equipment exhibits abnormal behavior. By adopting a so-called Total Productive Maintenance (TPM) philosophy, the size and workload of the maintenance department may well be reduced significantly, while at the same time significant improvements can be achieved in overall system effectiveness (Nakajima 1988). In this thesis, however, we will restrict ourselves to operational rather than organizational aspects of maintenance management.

1.3.2 Maintenance concept

Although many steps can be undertaken to maintain or improve the performance of a production system, only a few of them are normally considered to be the responsibility of the maintenance department. The most fundamental decision problems that are faced by maintenance management are:

- which items are to be maintained?
- what kind of maintenance must be conducted?
- when should these activities take place?

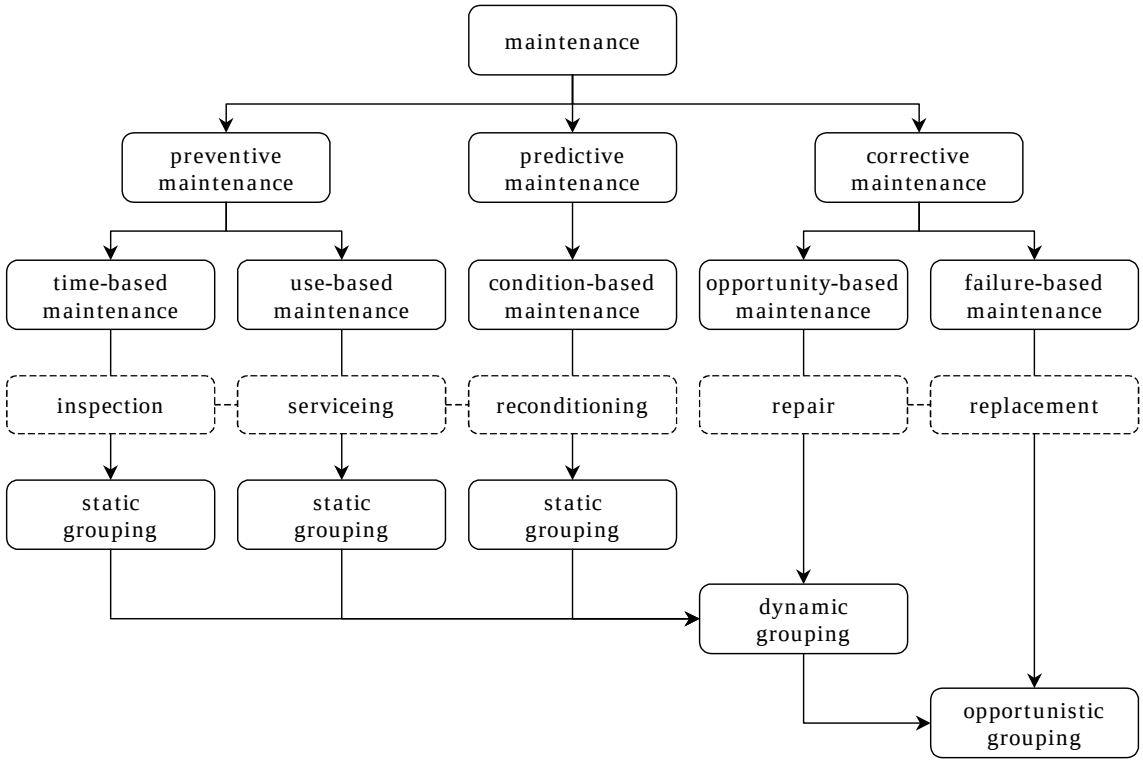


Figure 1.2: Maintenance initiators, activities, and packages: an overall perspective.

In general, these decisions are expressed in terms of an overall maintenance concept, prescribing which maintenance activities must be carried out at which times, and under which conditions (Gits 1984). In view of such a maintenance concept, a clear distinction is usually made between maintenance initiators, maintenance activities, and maintenance packages (see Figure 1.2). Within safety and legislative restrictions, a maintenance concept should be based upon an overall attempt to minimize both direct maintenance costs (e.g. labour, materials), and indirect maintenance costs (e.g. production loss, deterioration costs). In practice, estimation of indirect maintenance costs is usually very difficult. Nevertheless, they are often much larger than direct maintenance costs (Pintelon and Gelders 1992). In this thesis, we will account for both direct and indirect maintenance costs in formulating optimal maintenance strategies for unreliable production systems.

1.3.3 Maintenance initiators

As a starting point, there has to be some kind of control mechanism with which the need for maintenance is initiated. In practice, a categorization is usually made into preventive maintenance (before failure), predictive maintenance (just before fail-

ure), and corrective maintenance (upon or after failure). More specifically, the most commonly applied maintenance initiators are:

- time-based maintenance,
- use-based maintenance,
- condition-based maintenance,
- opportunity-based maintenance,
- failure-based maintenance.

Preventive maintenance is planned and performed before failure, and is either time-based (e.g. on a weekly basis), use-based (e.g. based on running hours), or a combination of both. Predictive maintenance aims at the initiation of preventive maintenance just before failure, and is mostly condition-based (e.g. if too much vibration is observed). Finally, corrective maintenance is performed upon or after failure, and is either opportunity-based (not urgent) or failure-based (urgent). Since condition-based maintenance usually requires a deeper understanding of failure causes and predictors (e.g. crack growth, vibration analysis, thermography) of the production equipment under consideration, we will mainly restrict ourselves to time-based, use-based, opportunity-based and failure-based maintenance strategies in this thesis.

1.3.4 Maintenance activitities

Excluding the influence of technological improvements in equipment design and layout (e.g. modifications), discussion of which is not appropriate throughout this thesis, the most basic maintenance activities can be classified as follows:

- inspection,
- serviceing,
- reconditioning,
- repair,
- replacement.

As a starting point, most items will be inspected regularly, in order to detect any signs of reduced effectiveness and/or impending failure. Additionally, items will normally be serviced at regular intervals (e.g. readjusted, lubricated, cleaned) in order to ensure continued effective operation in the future. Moreover, reconditioning activities will often be carried out in order to sustain satisfactory operation of items or equipment before they fail. Upon failure, repairs will normally be required to restore the equipment into satisfactory operation. Finally, replacement of items and equipment will occur when they are no longer capable of proper functioning, or are beyond economic repair.

1.3.5 Maintenance packages

Since maintenance activities often require one or more preparatory set-up activities (e.g. crew travelling, equipment rental, dismantling), there is a perspective of significant gains if they can be carried out simultaneously (maintenance grouping). In this respect, we must at least distinguish between the following grouping possibilities:

- static grouping,
- dynamic grouping,
- opportunistic grouping.

In the long term, planned preventive maintenance activities are usually combined into so-called preventive maintenance packages, each of which is treated as a single maintenance activity in an operational planning phase (static grouping). In the medium term, planned preventive maintenance activities can be combined with each other, and with plannable corrective maintenance activities as well (dynamic grouping). In the short term, unplanned corrective maintenance activities can be combined with planned preventive and/or corrective maintenance activities (opportunistic grouping). Although each grouping strategy takes place at a different planning level, their mutual objective is to improve efficiency in terms of reducing set-up times and costs in an operational planning phase. In this thesis, we will mainly restrict ourselves to opportunities for static grouping.

1.4 Maintenance modelling

Simply stated, the overall objective of maintenance optimization models at a tactical planning level, is to determine the frequency and timing of preventive maintenance

activities, in order to arrive at an optimal balance between the costs and benefits of preventive and corrective maintenance. In this respect, it is not only direct maintenance costs (e.g. salaries, spare parts, tools), but also indirect maintenance costs (e.g. poor quality, delay penalties, safety stocks) that counts. In practice, estimation of indirect maintenance costs may be very difficult. Nevertheless, they are often much larger than direct maintenance costs (Pintelon et al. 1997).

Without loss of generality, and in order to simplify our discussion of existing models, we will assume that a production system is always inspected before it is maintained preventively. By doing this, it is the main responsibility of the maintenance department to determine for each system (i) what to inspect, (ii) when to inspect, (iii) when to maintain, and (iv) how to maintain. Traditionally, the majority of maintenance optimization models has been concentrating on the second and third problem areas, whereas the remaining decisions are usually left to specialized maintenance engineers. In other words, the problem areas that are usually addressed by maintenance optimization models can be stated as follows:

- when should we inspect the system?
- on the basis of these inspection results, when should we maintain the system?

As a starting point, a clear distinction must be made between mathematical models for single unit systems, and mathematical models for multiple unit systems, since the latter implies the existence of economies of scale in conducting maintenance activities simultaneously. Within each of these classes, it is in our view essential to distinguish between three fundamental types of maintenance optimization models, as will be stated more explicitly in the following sections.

1.4.1 Single unit systems without economic dependence

The general philosophy of most maintenance optimization models for single unit systems, is to decide at each feasible moment whether it is cost-effective to carry out preventive maintenance now, or to postpone it to the next feasible moment, e.g. see Berg (1980) and Frenk et al. (1997). As a consequence, the main differences between these models originate from their interpretation of feasible moments, or equivalently the mechanism with which preventive maintenance is, or can be activated. In this respect, a clear distinction must be made between continuous review models, periodic review models, and opportunistic review models. The reader is referred to McCall (1965), Pierskalla and Voelker (1976), Sherif and Smith (1981), and Valdez-Flores and Feldman (1989) for a more comprehensive review on existing literature. Here, we will

only mention some important references, and discuss the scientific contributions of this thesis in some more detail.

Continuous review models

In continuous review models, it is assumed that the condition of the system can be monitored continuously. As a consequence, preventive maintenance is usually of a predictive, condition-based nature. According to Niebel (1994), there are five basic techniques typically used in condition monitoring: vibration monitoring, process parameter monitoring, thermography, tribology and visual inspection. Continuous monitoring of those parameters that allow the accurate prediction of failure will permit precise scheduling of repairs without the costs of emergency downtime. Mathematical models in this area derive their value from finding the parameters, and corresponding treshold values, with which the occurence of failures can be predicted accurately. The reader is referred to e.g. Barron (1996) for an introduction into the practice, methods and applications of condition monitoring techniques.

Periodic review models

In periodic review models, it is assumed that the condition of the system cannot be monitored continuously, as is the case in continuous review models, but only through periodic inspection at fixed costs. In these models, inspections are usually carried out at regular intervals, and are either time-based or use-based. In general, use-based maintenance policies outperform time-based maintenance policies in view of efficiency. On the other hand, time-based maintenance policies do have the advantage that one does not have to keep track of individual component ages, as a result of which they can easily be implemented and executed in a practical context. Mathematical models in this area are usually concerned with finding the optimal maintenance interval, either time-based or use-based, in order to arrive at an optimal balance between the costs of preventive and corrective maintenance. Well-known maintenance models of the use-based type are the age replacement and minimal repair model (Barlow and Hunter 1960). Classical examples of time-based maintenance models are the block replacement model (Barlow and Proschan 1965), the modified block replacement model (Berg and Epstein 1976), the standard inspection model (Barlow et al. 1963), and the delay time model (Christer 1982).

Opportunistic review models

In opportunistic review models, it is assumed that inspections cannot be carried out at any time, as is the case in periodic review models, but only at so-called maintenance opportunities. The underlying observation behind these models is that in many practical situations, preventive maintenance on non-critical units is delayed to some moment in time where the unit is not required for production. In general, such opportunities may arise due to e.g. random breakdowns and/or withdrawn production orders. Because of the random occurence of opportunities, and because of their sometimes restricted duration, traditional maintenance models fail to make effective use of them. Mathematical models in this area are primarily used to determine whether a maintenance activity must be conducted at a given opportunity, or whether it must be postponed to the next one, e.g. see Berg (1984), Dekker and Smeitink (1991), and Dekker and Dijkstra (1992).

Contribution of this thesis

In this thesis, we will restrict ourselves to periodic and opportunistic review models. Traditionally, maintenance optimization models for single unit systems have been focussed on minimizing the long run average times and/or costs associated with preventive and corrective maintenance. Moreover, the initiation of preventive maintenance is often either time-based, or use-based, or opportunity-based, but nothing in between. Simply stated, the models presented in this thesis generalize previous work in the latter respect, and also by taking into account interactions with production in terms of variability measures. In chapter 4, we study the effect of preventive maintenance policies on the interval availability distribution of an unreliable production system. In chapter 5, we present a combined periodic/opportunistic review model, in which preventive maintenance is carried out at the best opportunity during a predefined interval. Although these philosophies are common sense in practice, they certainly are an underexposed point of view in existing literature.

1.4.2 Multiple unit systems with economic dependence

The justification of most maintenance optimization models for multiple unit systems, is a potential of reductions in set-up costs and/or times if maintenance activities are carried out simultaneously (maintenance grouping). As we explained in the previous section, mathematical models in this area can be categorized into static grouping, dynamic grouping, and opportunistic grouping strategies. Although each grouping

strategy takes place at a different planning level, their mutual objective is to improve efficiency in terms of reducing set-up times and costs in an operational planning phase. The reader is referred to Cho and Parlar (1991) and Dekker et al. (1997) for an extensive and up-to-date literature review on maintenance models for multi-unit systems with economic dependence. Here, we will restrict ourselves to some important references. Moreover, we discuss the contributions of this thesis to existing literature.

Static grouping models

Static grouping refers to the combination of planned preventive maintenance activities in a strategical planning phase. In this respect, a clear distinction must be made between direct and indirect grouping models. In direct grouping models, the collection of preventive maintenance activities is partitioned into several maintenance packages, each of which is executed at an interval that is optimal for that package. In indirect grouping models, maintenance packages are not determined in advance, but are formed indirectly whenever the maintenance of different units coincides. To achieve this, each maintenance activity is carried out at an integer multiple of a certain basis interval. Basically, static grouping models attempt to find the optimal balance between the costs of deviating from the optimal preventive maintenance intervals for individual units, and the benefits of combining preventive maintenance activities on different units. Typical examples of static grouping models can be found in e.g. Gertsbakh (1977), Goyal and Kusy (1985), Goyal and Gunasekaran (1992), and Wildeman (1996).

Dynamic grouping models

Dynamic grouping refers to the combination of planned preventive maintenance activities with each other, and/or with plannable corrective maintenance activities, in a tactical planning phase. Of course, the latter is only possible if the repair of failed units can be postponed to a more suitable moment in time, e.g. because standby units are available, or the unit does not affect the system as a whole. The main difficulty of dynamic grouping models is that the failure of a unit cannot be predicted in advance. Therefore, dynamic grouping models often make use of a so-called rolling horizon approach. More specifically, they use a finite horizon in order to arrive at a sequence of decisions, but only implement the first one. Basically, mathematical models for dynamic grouping derive their value from finding an optimal balance between the costs of postponing corrective maintenance activities, and the benefits of combining them with other preventive and/or corrective maintenance activities. Typical

examples of dynamic grouping models are presented in e.g. Assaf and Shanthikumar (1987), Ritchken and Wilson (1990), Jansen and Van der Duyn Schouten (1995), and Wildeman et al. (1997).

Opportunistic grouping models

Opportunistic grouping refers to the combination of planned maintenance activities with unplanned maintenance activities in an operational planning phase. In these models, the failure of a particular unit is used as an opportunity for planned maintenance on other units. In practice, this means that opportunistic maintenance grouping is difficult to manage, since it affects the plannable nature of preventive maintenance. Nevertheless, if all the preparations needed for preventive maintenance have been made in advance, it certainly is an effective method to reduce set-up costs and times in an operational planning phase. Basically, mathematical models for opportunistic grouping attempt to find an optimal balance between the costs of advancing planned maintenance activities, and the benefits of combining them with other unplanned maintenance activities. Typical examples of opportunistic grouping models can be found in e.g. Haurie and L'Ecuyer (1982), Ozekici (1988), Van der Duyn Schouten and Vanneste (1990), Van der Duyn Schouten and Vanneste (1993), Dekker and Smeitink (1994), and Wijnmalen and Hontelez (1997).

Contribution of this thesis

In this thesis, we will mainly be concerned with static grouping or so-called clustering models. Traditionally, clustering models have accounted for the economies of scale in carrying out maintenance jobs simultaneously, by assuming that a fixed set-up cost is incurred at each occasion for preventive and/or corrective maintenance. In this thesis, we will present a new, much richer and more powerful modelling framework, which allows for multiple interrelated set-up activities. In chapter 2, this framework is applied to a direct clustering model, in which a collection of frequency-constrained maintenance jobs must be subdivided into several frequency-constrained maintenance packages, and our objective is to minimize preventive maintenance costs per unit of time. In chapter 3, we consider an indirect clustering model, with more general frequency-dependent costs instead of frequency constraints for each component. Here, each component is maintained preventively at an integer multiple of a certain basis interval, which is the same for all components. Our approach generalizes previous work in the sense that considerably more degrees of freedom are taken into account.

1.5 Overview of this thesis

Let us now briefly discuss the contents of this thesis in some more detail. In chapters 2 and 3, we present a modelling framework with which the costs and times associated with preventive and corrective maintenance can be modelled to a proper level of detail, and for a large class of production systems. The underlying observation behind this modelling framework is that almost all production systems can be decomposed hierarchically into a tree-like structure of set-up activities and components, in which each component corresponds to exactly one set-up activity. In line with this, creating an occasion for preventive maintenance on one of these components requires a collection of preparatory set-up activities to be carried out in advance, with corresponding set-up times and/or costs. Since different components may require one or more identical or shared set-up activities, there is a perspective of significant gains if preventive maintenance activities are carried out simultaneously. By assuming an additive cost structure, the opportunities for static grouping are further exploited in chapters 2 and 3 respectively. These chapters are partially based on ideas that were first presented in Van Dijkhuizen (1995).

Chapter 2 considers a direct clustering problem for a multi-setup and multi-component production system with frequency-constrained preventive maintenance jobs, which must be carried out with prescribed or higher frequencies. This approach is particularly useful if these frequencies are restricted by law (e.g. aircraft maintenance), or historical data are simply not available (e.g. new equipment). Our objective in this chapter is to find a partitioning of preventive maintenance jobs into preventive maintenance packages, in such a way that overall preventive maintenance costs per unit of time are minimized. To this end, a clear distinction is made between production systems with a single set-up activity (common set-ups), and production systems with multiple set-up activities (shared set-ups), since the former requires much simpler solution techniques than the latter. A preliminary version of this chapter has been published in Van Dijkhuizen and Van Harten (1997a).

In chapter 3, our modelling framework is further developed into an indirect grouping problem, in which each component is maintained preventively at integer multiples of a certain basis interval, and corrective maintenance is carried out in between whenever necessary. Within this setting, our objective is to determine a repetitive maintenance cycle which minimizes the average maintenance costs per unit of time in the long run. To this end, detailed information about the failure behavior of each component is assumed to be available. More specifically, the frequency constraints of chapter 2 are replaced with frequency-dependent costs for each maintenance job. Our

approach generalizes previous work in the sense that it allows for multiple interrelated set-up activities and components, and that considerably more degrees of freedom are taken into account. The majority of this chapter is based on results presented in Van Dijkhuizen and Van Harten (1997b).

Chapter 4 is concerned with the interval availability distribution of an unreliable production system, which is maintained preventively at regular intervals, and correctively upon failure. Within this setting, it is examined whether, and to which extent, optimal preventive maintenance policies would change if the guaranteed performance of a production system during a finite period of time (interval availability) would be preferred above its average performance in the long run (limiting availability). A general modelling framework is presented which allows for stochastic preventive and corrective maintenance times. In addition, explicit formulas are derived for a production system with deterministic preventive maintenance times, and Gamma-distributed corrective maintenance times. Computational results indicate that frequent and short service interruptions are to be preferred above infrequent and long ones, all other things being equal. The results of this chapter are primarily based on Van Dijkhuizen and Van der Heijden (1998).

Chapter 5 is concerned with the potential benefits of building in some flexibility concerning the starting time of preventive maintenance. The underlying observation behind this approach is that the initiation of preventive maintenance should be based on the technical state as well as the operating state of a production system, and that the latter is often subject to fluctuations in time. Although this is a widespread common sense in practice, it certainly is an underexposed point of view in existing literature. Therefore, a two-stage maintenance policy is considered, which - in a first stage - uses the technical state of the production system to determine a finite interval during which preventive maintenance must be carried out, and - in a second stage - uses the operating state of the production system to determine the optimal starting time within that interval. Computational results indicate that significant savings can be obtained in comparison with classical maintenance policies. These chapters have been published in adapted form in Van Dijkhuizen and Van Harten (1998a) and Van Dijkhuizen and Van Harten (1998b).

In chapter 6, we will present the results of a case study that was carried out at the Line Maintenance department of KLM Royal Dutch Airlines. This department is responsible for the inspection, maintenance and repair of aircrafts during their stay at Schiphol Airport, as well as the assignment of aircrafts to flights in KLM's timetable. A decision support system has been developed which should eventually assist maintenance managers in determining how many maintenance slots of which

type should be available in the timetable, and how many maintenance engineers of which type should be assigned to these slots. This chapter is partially based on ideas presented in Van Dijkhuizen (1997).

Finally, chapter 7 summarizes the ideas and models presented in this thesis, and clears the way towards a decision support system for coordinated planning and scheduling of production and maintenance. We conclude that preventive maintenance frequencies should be determined in a strategical planning phase, whereas tactical and operational decision making should be supported with relatively simple, and rather straightforward control mechanisms. Therefore, a clear distinction is made between long term, medium term, and short term maintenance planning. To conclude this thesis, we briefly discuss these control mechanisms at each planning level, in view of the possible interactions with production, and indicate several interesting opportunities for further research as well.

Chapter 2

Optimal clustering of frequency-constrained maintenance jobs with shared set-ups

Since maintenance jobs often require either one, or a collection of preparatory set-up activities, joint execution or clustering of maintenance jobs is often seen as a powerful instrument to reduce shut-down costs. In this chapter, we consider a clustering problem for frequency-constrained maintenance jobs, i.e. maintenance jobs that must be carried out with prescribed (or higher) frequencies. As a starting point, several strong dominance rules are provided for the clustering of maintenance jobs with identical, so-called common set-ups. Subsequently, these dominance rules are used in an efficient dynamic programming algorithm, which solves this problem in polynomial time. For the clustering of maintenance jobs with partially identical, so-called shared set-ups, similar but less strong dominance rules are derived. Nevertheless, a dynamic programming algorithm and a mixed integer linear program, as well as two surprisingly well-performing heuristics, can be formulated to solve this problem.

2.1 Introduction

Many preventive maintenance jobs (inspections, replacements) of production systems require shut-down of the units involved. If these units are used continuously, as is the case in process industry, shut-downs can be very costly and management will try to minimize their duration and frequency. Since maintenance jobs often share one or more preparatory set-up activities and/or costs (e.g. crew travelling, dismantling, equipment rental, carry in at a repair facility), there is a perspective of significant

gains if they can be carried out simultaneously (maintenance grouping). In the past few decades, a growing interest can be observed in the modelling and optimization of preventive and corrective maintenance planning, where correlation between various jobs is essential in view of set-up avoidance. Most of these models derive their value from the economies of scale in carrying out maintenance jobs simultaneously, e.g. see Cho and Parlar (1991) and Dekker et al. (1997) for an extensive and up-to-date review of existing literature. Here, we will only mention some important references. As in the previous chapter, we make a clear distinction between static, dynamic and opportunistic grouping models.

First of all, static grouping refers to the coordination of planned preventive maintenance jobs in a strategical planning phase, e.g. see Gertsbakh (1977), Goyal and Kusy (1985), Goyal and Gunasekaran (1992), and Wildeman (1996). In a similar way, dynamic grouping models aim at the simultaneous execution of planned preventive maintenance jobs and plannable corrective maintenance jobs in a tactical planning phase, e.g. see Assaf and Shanthikumar (1987), Ritchken and Wilson (1990), Jansen and Van der Duyn Schouten (1995), and Wildeman, Dekker, and Smit (1997). Finally, opportunistic grouping comes down to the combination of planned with unplanned maintenance jobs in an operational planning phase, e.g. see Haurie and L'Ecuyer (1982), Ozekici (1988), Van der Duyn Schouten and Vanneste (1990), Van der Duyn Schouten and Vanneste (1993), Dekker and Smeitink (1994), and Wijnmalen and Hontelez (1997). Basically, the objective of maintenance grouping models is to find an optimal balance between the costs of deviating from the optimal interval for individual maintenance jobs, and the benefits in terms of set-up avoidance by combining maintenance jobs. In this chapter, we will focus on static grouping, or **long term clustering possibilities** for preventive maintenance jobs.

In this respect, a clear distinction must be made between direct and indirect clustering techniques. In direct clustering models, the collection of preventive maintenance activities is partitioned into several maintenance packages, each of which is executed at an interval that is optimal for that package. In indirect clustering models, these maintenance packages are not determined in advance, but are formed indirectly whenever the occurence of different maintenance jobs coincides. To achieve this, each maintenance job is carried out at an integer multiple of a certain basis interval. In general, indirect clustering techniques always outperform direct clustering techniques, at least from a cost efficiency point of view (Van Eijs, Heuts, and Kleinen 1992). Nevertheless, there may be some other e.g. administrative reasons which emphasize on the use of predefined maintenance packages. In this chapter, we will focus on direct clustering techniques.

From a theoretical point of view, the clustering problem can be formulated as a classical set-partitioning problem, and as such is $\mathcal{NP}$-hard (Garey and Johnson 1979). Consequently, an optimal clustering can be found for only a relatively small number of maintenance jobs, unless a very special structure is assumed. An example of this type is presented in this chapter, where we restrict ourselves to **frequency-constrained maintenance jobs**. More specifically, we consider maintenance jobs that must be carried out at fixed intervals with prescribed or higher frequencies (e.g. at least once per month). Although it is very well possible that these so-called limitative frequencies are determined with the use of mathematical models, they might also be based on expert opinions, safety restrictions and/or legislation. In general, the use of frequency constraints requires no explicit data on failure statistics whatsoever, and therefore enables us to integrate both qualitative and quantitative decisions in one and the same modelling framework.

Pioneering work within the field of frequency-constrained maintenance jobs was carried out by Gits (1987), who considered a clustering problem for maintenance jobs with identical, so-called common set-up activities. The latter implies that creating an occasion for preventive maintenance on one or more components requires a fixed set-up cost, irrespective of how many and which components are maintained. Although this might be an interesting approach from a theoretical point of view, it is obvious that nowadays production systems are usually much more complicated. In our opinion, preventive maintenance models should at least account for multiple set-up activities and components, allowing different set-up costs for different components, or groups of components. On the other hand, it seems virtually impossible to support a separate, independent data structure for each possible group of components, that could arise in the most general situation.

In order to arrive at a compromise, we developed a powerful modelling framework, in which a variety of complex set-up structures can be modelled to a proper level of detail. More specifically, we assume that the collection of set-up activities can be ordered hierarchically into a tree-like structure, in which each maintenance job can be associated with exactly one set-up activity (see Figure 2.1). Considering this tree-like structure, it is now immediately clear that some maintenance jobs may not share all set-up activities, but only a subset of them. Obviously, these possibilities for **shared set-up activities** provide a richer and more realistic modelling framework in comparison with the requirement of completely coinciding paths, as is the case with common set-ups.

To a certain extent, the above-mentioned concept of shared set-up activities originates from Gertsbakh (1972), who developed a somewhat similar but less powerful

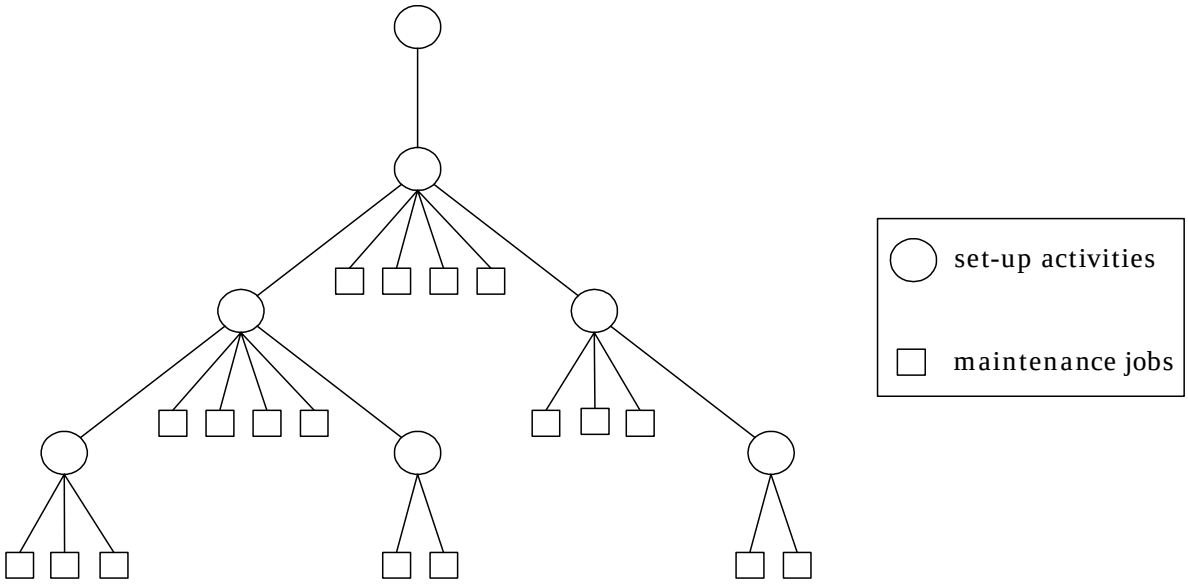
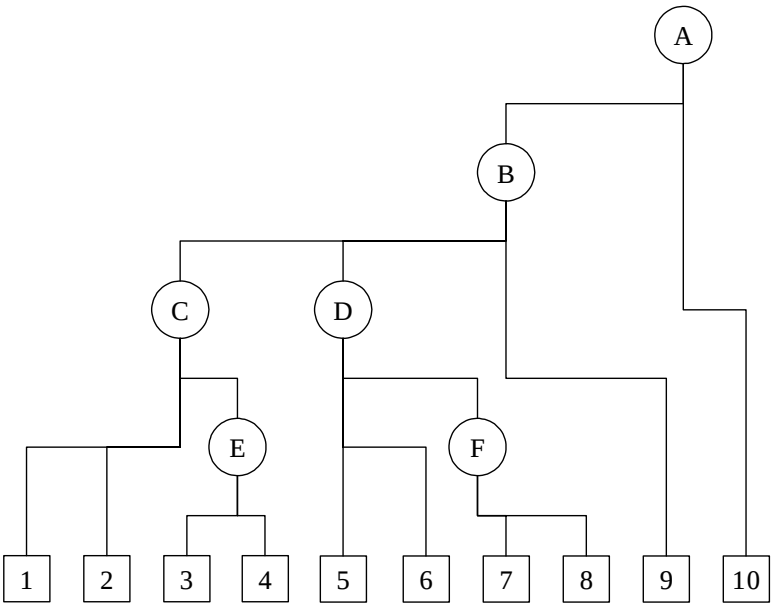


Figure 2.1: Tree-like structure of set-up activities and maintenance jobs (example).

modelling framework, in which maintenance jobs can only be defined at the lowest-level set-up activities. Practical examples of shared, but not common set-ups can be found in various areas, e.g. in aircraft maintenance, maintenance of nuclear power plants, off-shore maintenance, and even tramcar maintenance (see Figure 2.2). Summarizing, the notion of shared but not common set-up activities seems to be a common sense in practice, but at the same time an underexposed point of view in existing literature.

The outline of this chapter is as follows. In section 2.2, a mathematical formulation of the clustering problem is given, and the complexity of this problem is briefly discussed. In section 2.3, the clustering problem with common set-ups is considered. Several dominance rules are provided, and an efficient dynamic programming algorithm is developed, which solves this problem in polynomial time. In section 2.4, the clustering problem with shared set-ups is considered, and similar but less strong dominance rules are derived. Nevertheless, a dynamic programming algorithm and a mixed integer linear program can be formulated, with which this clustering problem can also be solved to optimality. Subsequently, two efficient heuristics are presented in section 2.5, and their absolute as well as relative performance is further investigated in section 2.6. Finally, some conclusions are summarized in section 2.7, and several possibilities for further research are discussed.



Removing and fitting body to truck assembly	120
Removing and fitting motors to truck	300
Opening motor and removing armature plus reassembling	50
Dismantling motor housing and removing interpole and field coils plus reassembling	50
Reconditioning armature	900
Reconditioning field coil	120
Reconditioning interpole coil	100
New roller bearing	350
Major truck overhaul	5000
Minor truck overhaul	1900
Body overhaul and painting	3000

Figure 2.2: Multiple interrelated set-up and maintenance activities with associated costs for a tramcar (Sculli and Suraweera 1979).

2.2 General approach

In this section, the underlying assumptions of our modelling framework are stated more explicity. Furthermore, a proper problem definition and a mathematical formulation of the clustering problem are given. Moreover, the complexity of this problem in terms of the number of possible clusterings in relation to the number of maintenance jobs is discussed.

2.2.1 Problem definition

As a starting point, we assume that fixed costs are incurred for each set-up activity, and for each maintenance job. In general, these costs can be categorized into direct

maintenance costs (e.g. salaries, tools, materials), and indirect maintenance costs (e.g. production loss, delay penalties). Given a limitative frequency or frequency constraint for each maintenance job, the clustering problem is now concerned with the partitioning (clustering) of maintenance jobs into maintenance packages (clusters), in such a way that preventive maintenance costs per unit of time are minimized in the long run. Note that the possible reductions in equipment failures and associated corrective maintenance costs, as a positive side-effect of clustering, are not contained in our analysis. If so, clustering of preventive maintenance jobs would become even more profitable.

Within our modelling framework, we assume that the costs of a cluster can be computed from the costs of the individual set-up activities and maintenance jobs, and that the costs of a clustering can be computed from the costs of the individual clusters. In other words, we use an overall additive cost structure, as will be stated more explicitly in the following section. From a practical point of view, this means that (i) parallel execution of maintenance jobs within a cluster, and (ii) simultaneous execution of clusters within a clustering (e.g. in an operational planning phase) are not allowed, or do not lead to cost reductions. Obviously, other assumptions would lead to other interesting versions of the clustering problem. In fact, an illustrative example will be presented in the following chapter.

In line with the above, the collection of set-up activities and frequency-constrained maintenance jobs is now converted into a so-called **maintenance tree**. The root of this maintenance tree corresponds to the production system in operating condition. Moreover, maintenance jobs are represented by the leafs, and set-up activities by the remaining nodes of the tree. As a consequence, each maintenance job can be associated with exactly one parental set-up activity, and the communality of set-up activities is determined by the joint part of the paths connecting the nodes to the root of the tree (see Figure 2.3). These are the basic rules for the conversion, further details are discussed below.

In general, this tree-like structure of set-up activities and maintenance jobs relates to the hierarchical decomposition of a production system into its several subsystems. This does not necessarily imply, however, that other possibilities do not exist. As an illustrative example, consider a melting furnace which is subject to several periodic preventive maintenance jobs. Due to safety restrictions, different maintenance jobs require different furnace temperatures. The furnace has to be cooled down to the required temperature before a specific maintenance job can be carried out. If we associate a set-up activity with each of the required temperatures, then the set of different temperatures also reflects a shared set-up structure.

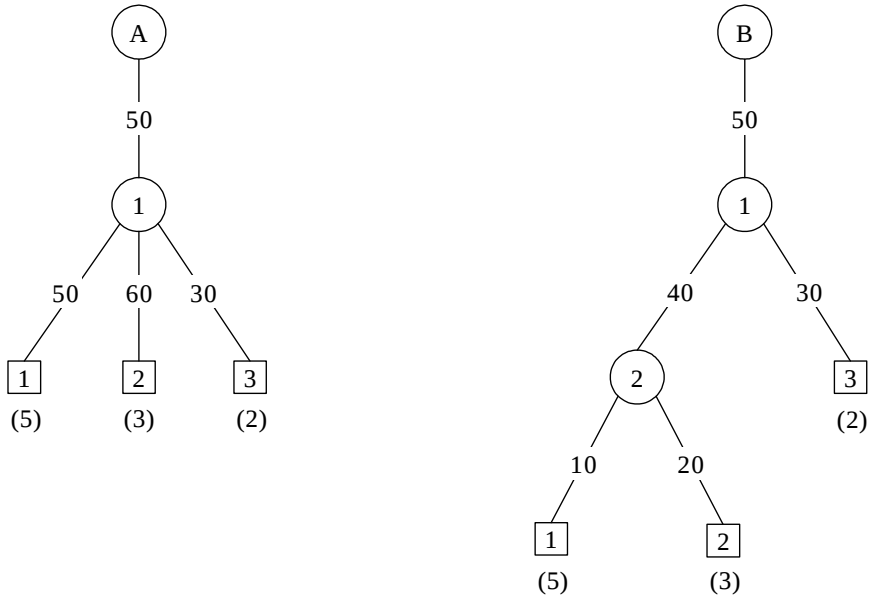


Figure 2.3: Examples of a maintenance tree with (a) common and (b) shared set-up activities: set-up and maintenance costs are shown at the arcs, limitative frequencies in brackets at the corresponding nodes.

2.2.2 Mathematical formulation

Consider a collection of preparatory set-up activities $\mathcal{I} = \{1, ..., m\}$ and frequency-constrained maintenance jobs $\mathcal{J} = \{1, ..., n\}$, and let $I_j \subseteq \mathcal{I}$ denote the set-up activities $i \in \mathcal{I}$ required for maintenance job $j \in \mathcal{J}$ (e.g. $I_2 = \{1, 2\}$ in Figure 2.3b). Furthermore, define $s_i > 0$ and $c_j > 0$ as the (expected) costs associated with set-up activity $i \in \mathcal{I}$ resp. maintenance job $j \in \mathcal{J}$, and let $f_j > 0$ denote the limitative frequency of maintenance job $j \in \mathcal{J}$.

A **cluster** of maintenance jobs is defined as a subset $U \subseteq \mathcal{J}$. Similar to the definitions above, let $f(U) > 0$ denote the limitative frequency, $s(U) > 0$ the (expected) set-up costs and $c(U) > 0$ the (expected) maintenance costs associated with a cluster $U \subseteq \mathcal{J}$. Then the following expressions for $f(U)$, $s(U)$ and $c(U)$ can be derived:

$$f(U) = \max_{j \in U} f_j,$$

$$s(U) = \sum_{i \in \bigcup\limits_{j \in U} I_j} s_i,$$

$$c(U) = \sum_{j \in U} c_j.$$

In a similar way, the (expected) costs per unit of time $\lambda(U)$, associated with a cluster $U \subseteq \mathcal{J}$, can be defined rather straightforwardly as follows:

$$\lambda(U) = f(U) \cdot \{s(U) + c(U)\}.$$

A **clustering** of maintenance jobs is defined as a partitioning Ω of $\mathcal{J}$. Since our objective is to minimize the total (expected) costs per unit time, we need to determine a clustering Ω^* for which the following expression is minimized:

$$\Lambda(\Omega) = \sum_{U \in \Omega} \lambda(U) = \sum_{U \in \Omega} f(U) \cdot \{s(U) + c(U)\}.$$

2.2.3 Further notation and assumptions

As a starting point, we define $J_i = \{j \in \mathcal{J} \mid i \in I_j\}$ and $J_i^* \subseteq J_i$ as the collection of maintenance jobs $j \in \mathcal{J}$ that require resp. are attached to set-up activity $i \in \mathcal{I}$. Moreover, we denote with $S_i \subset \mathcal{I}$ the successors of set-up activity $i \in \mathcal{I}$, and with $\mathcal{K} = \{f_j \mid j \in \mathcal{J}\}$ the entire set of different limitative maintenance frequencies. Finally, we define $K_j = \{k \in \mathcal{K} \mid k \geq f_j\}$ and $\mathcal{K}_i = \bigcup_{j \in J_i} K_j$ as the set of feasible maintenance frequencies for maintenance job $j \in \mathcal{J}$ resp. set-up activity $i \in \mathcal{I}$. As an illustrative example, we consider the maintenance tree of Figure 2.3b. This yields:

$$\begin{gathered} I_1 = I_2 = \{1,2\},\ I_3 = \{1\}, \\ J_1 = \{1,2,3\},\ J_2 = \{1,2\}, \\ J_1^* = \{3\},\ J_2^* = \{1,2\}, \\ S_1 = \{2\},\ S_2 = \emptyset, \\ \mathcal{K} = \mathcal{K}_1 = \{5,3,2\},\ \mathcal{K}_2 = \{5,3\}, \\ K_1 = \{5\},\ K_2 = \{5,3\},\ K_3 = \{5,3,2\}. \end{gathered}$$

For notational convenience, and without loss of generality, we assume that $i' > i$ for all $i \in \mathcal{I}$ and $i' \in S_i$ in the sequel. Moreover, and in line with the above, we assume that $I_1 \cap ... \cap I_n = \{1\}$, i.e. there exists exactly one common set-up activity $i = 1$. To see this, $I_1 \cap ... \cap I_n = \emptyset$ implies that the clustering problem can be decomposed into two or more (smaller) clustering problems, which can be treated and solved separately. In case $|I_1 \cap ... \cap I_n| = k > 1$, the common set-up activities $i \in \{1, ..., k\}$ can as well be combined into a single common set-up activity $i = 1$, with corresponding set-up costs $s_1 + ... + s_k > 0$. For similar reasons, we can assume that $J_i \neq \varnothing$ for all $i \in \mathcal{I}$, since obviously set-up activities with no components can as well be neglected.

Table 2.1: Total number of different clusterings $A(n)$ for a clustering problem with n maintenance jobs.

n	2	6	12	20	30
$A(n)$	2	203	$4.21 \cdot 10^6$	$5.17 \cdot 10^{13}$	$8.47 \cdot 10^{23}$

2.2.4 Complexity of the clustering problem

Let us now investigate the complexity of the clustering problem, by deriving an expression $A(n)$ for the number of different clusterings of $\{1, ..., n\}$. To this end, define $B(n, k)$ as the number of different clusterings of $\{1, ..., n\}$ into k clusters ($1 \leq k \leq n$). Then $A(n) = \sum_k B(n, k)$ by definition, whereas $B(n, 1) = B(n, n) = 1$ is almost trivial. Moreover, $B(n, k) = k \cdot B(n-1, k) + B(n-1, k-1)$ for $1 < k < n$. Some values of $A(n)$ for increasing values of n are presented in Table 2.1. Apparently, the search space of the clustering problem grows exponentially with the number of maintenance jobs n. In the following sections, we will show that the complexity of the clustering problem can be reduced significantly, by exploiting its special structure.

2.3 The clustering problem with common set-ups

In this section, the clustering problem for maintenance jobs with common set-ups is considered. First, an example is given, and several strong dominance rules are provided. With these dominance rules, an efficient dynamic programming algorithm is developed, which solves this problem in polynomial time. In the clustering problem with common set-ups, there is only one set-up activity ($m = 1$). Consequently, $s(U) = s_1$ for all $U \subseteq \mathcal{J}$, where $s_1 > 0$ represents the common set-up costs.

2.3.1 Example

Consider the clustering problem with common set-ups, as shown in Figure 2.3a. Then the costs $\lambda(U)$ for all possible clusters $U \subseteq \mathcal{J}$ are determined as follows:

$$\begin{aligned}
\lambda(\{1\}) &= 5 \cdot (50 + 50) = 500, \\
\lambda(\{2\}) &= 3 \cdot (50 + 60) = 330, \\
\lambda(\{3\}) &= 2 \cdot (50 + 30) = 160, \\
\lambda(\{1, 2\}) &= 5 \cdot (50 + 50 + 60) = 800, \\
\lambda(\{1, 3\}) &= 5 \cdot (50 + 50 + 30) = 650, \\
\lambda(\{2, 3\}) &= 3 \cdot (50 + 60 + 30) = 420, \\
\lambda(\{1, 2, 3\}) &= 5 \cdot (50 + 50 + 60 + 30) = 950.
\end{aligned}$$

In a similar way, the costs $\Lambda(\Omega)$ for all possible clusterings Ω of $\mathcal{J}$ are given by:

$$\begin{gathered}\Lambda(\{\{1\},\{2\},\{3\}\}) = 500 + 330 + 160 = 990,\\ \Lambda(\{\{1\},\{2,3\}\}) = 500 + 420 = 920,\\ \Lambda(\{\{2\},\{1,3\}\}) = 330 + 650 = 980,\\ \Lambda(\{\{3\},\{1,2\}\}) = 160 + 800 = 960,\\ \Lambda(\{\{1,2,3\}\}) = 950.\end{gathered}$$

Apparently, the optimal clustering is determined by $\Omega^* = \{\{1\},\{2,3\}\}$, with corresponding (minimal) costs $\Lambda(\Omega^*) = 920$.

2.3.2 Problem reduction

Let us now derive some dominance rules, with which optimal clusterings can be characterized, and as a result of which the complexity of the clustering problem can be reduced significantly.

Theorem 1 *Consider an optimal clustering Ω^* and let $Q_j \in \Omega^*$ denote the cluster corresponding to maintenance job $j \in \mathcal{J}$. Then the following must be satisfied:*

(i) $\quad \forall j,k \in \mathcal{J} : f(Q_j) = f(Q_k) \rightarrow Q_j = Q_k,$
(ii) $\quad \forall j,k \in \mathcal{J} : f_j \geq f_k \rightarrow f(Q_j) \geq f(Q_k),$
(iii) $\quad \forall j \in \mathcal{J} : f(Q_j) \leq f_j \cdot (s_1 + c_j)/c_j,$
(iv) $\quad \forall j,k \in \mathcal{J} : f_j = f_k \rightarrow Q_j = Q_k,$
(v) $\quad \forall j,k,l \in \mathcal{J} : f_j \geq f_l \geq f_k \ \wedge \ Q_j = Q_k \rightarrow Q_j = Q_l = Q_k.$

Proof. If Ω^* violates (i), then $f(Q_j) = f(Q_k)$ and $Q_j \neq Q_k$ for some $j,k \in \mathcal{J}$. In that case, combination of clusters Q_j and Q_k results in a clustering Ω with $\Lambda(\Omega) = \Lambda(\Omega^*) - f(Q_j) \cdot s_1 < \Lambda(\Omega^*)$, since $s_1 > 0$ by assumption. If Ω^* violates (ii), then $f_j \geq f_k$ and $f(Q_j) < f(Q_k)$ for some $j,k \in \mathcal{J}$. In that case, moving job k from cluster Q_k to Q_j results in a clustering Ω with $\Lambda(\Omega) \leq \Lambda(\Omega^*) - f(Q_k) \cdot c_k + f(Q_j) \cdot c_k < \Lambda(\Omega^*)$, since $f(Q_j) \geq f_j \geq f_k$ and $c_k > 0$ by assumption. If Ω^* violates (iii), then $f(Q_j) \cdot c_j > f_j \cdot (s_1 + c_j)$ for some $j \in \mathcal{J}$. In that case, removing job j from cluster Q_j and creating a new cluster $\{j\}$ results in a clustering Ω with $\Lambda(\Omega) \leq \Lambda(\Omega^*) - f(Q_j) \cdot c_j + f_j \cdot (s_1 + c_j) < \Lambda(\Omega^*)$. Finally, properties (iv) and (v) follow directly from properties (i) and (ii). $\square$

Let us now denote with $A(n,k)$ the total number of different clusterings for n maintenance jobs, which satisfy dominance rules (i) and (ii) and thus (iv) and (v) of Theorem 1, given that there are $k = |\mathcal{K}|$ different maintenance frequencies. Then it is

Table 2.2: Total number of different clusterings $A(n,k)$ for a clustering problem with n maintenance jobs and k maintenance frequencies.

n	2	6	12	20	30
k	2	3	4	5	6
$A(n)$	2	203	$4.21 \cdot 10^6$	$5.17 \cdot 10^{13}$	$8.47 \cdot 10^{23}$
$A(n,k)$	2	4	8	16	32

easily verified that $A(n,k) = 2^{k-1}$. Some values of $A(n,k)$ versus $A(n)$ for increasing values of n and k are given in Table 2.2. Clearly, the complexity of the clustering problem with common set-ups is reduced drastically, even if dominance rule (iii) of Theorem 1 is left out of consideration.

2.3.3 A dynamic programming algorithm

Using dominance rule (iv) of Theorem 1, we can assume -without loss of generality- that $f_1 > ... > f_n$, since maintenance jobs with identical frequencies are always contained in the same cluster. In other words, jobs j and k with $f_j = f_k$ can as well be replaced by a single job $l = \{j\} \cup \{k\}$ with $f_l = f_j = f_k$ and $c_l = c_j + c_k$. As a starting point of our analysis, let us now denote with $\hat{f}_j > 0$ the maximal frequency of maintenance job $j \in \mathcal{J}$, according to dominance rule (iii) of Theorem 1:

$$\hat{f}_j = \left\lfloor f_j \cdot \frac{s_1 + c_j}{c_j} \right\rfloor .$$

Furthermore, let $g(k)$ denote the minimal costs for clustering the first k maintenance jobs ($1 \leq k \leq n$), and define $g(0) = 0$ for notational convenience. Using dominance rules (iii) and (v) of Theorem 1, $g(k)$ can now be determined recursively by means of the following dynamic programming equation:

$$g(k) = \min_{j \leq k : f_j \leq \hat{f}_k} \{ g(j-1) + f_j \cdot (s_1 + c_j + ... + c_k) \} .$$

In general, the clustering problem may have alternative optimal solutions. Nevertheless, the above-mentioned dynamic programming algorithm requires $\frac{1}{2} \cdot n \cdot (n+1)$ computations in the worst case (i.e. if $\hat{f}_j \geq f_1$ for all $j \in \mathcal{J}$), and therefore is an $O(n^2)$ algorithm.

2.3.4 Example (continued)

Consider the example of Figure 2.3a, for which it can easily be verified that $\hat{f}_1 = \left\lfloor 5 \cdot \frac{50+50}{50} \right\rfloor = 10$, $\hat{f}_2 = \left\lfloor 3 \cdot \frac{50+60}{60} \right\rfloor = 5$, and $\hat{f}_3 = \left\lfloor 2 \cdot \frac{50+30}{30} \right\rfloor = 5$. According to

these maximal frequencies, none of the possible clusters can be discarded in advance. Hence, the dynamic programming algorithm results in:

$$\begin{aligned}
g(1) &= f_1 \cdot (s_1 + c_1) \\
&= 5 \cdot (50 + 50) = 500, \\
g(2) &= \min\{f_1 \cdot (s_1 + c_1 + c_2), g(1) + f_2 \cdot (s_1 + c_2)\} \\
&= \min\{5 \cdot (50 + 50 + 60), 500 + 3 \cdot (50 + 60)\} \\
&= \min\{\underline{800}, 830\} = 800, \\
g(3) &= \min\{f_1 \cdot (s_1 + c_1 + c_2 + c_3), g(1) + f_2 \cdot (s_1 + c_2 + c_3), g(2) + f_3 \cdot (s_1 + c_3)\} \\
&= \min\{5 \cdot (50 + 50 + 60 + 30), 500 + 3 \cdot (50 + 60 + 30), 800 + 2 \cdot (50 + 30)\} \\
&= \min\{950, \underline{920}, 960\} = 920.
\end{aligned}$$

Hence, the optimal solution is $\Omega^* = \{\{1\}, \{2, 3\}\}$, with corresponding (minimal) costs $\Lambda(\Omega^*) = 920$.

2.4 The clustering problem with shared set-ups

In this section, the clustering problem for maintenance jobs with shared set-ups is considered. First of all, a numerical example is given, and several dominance rules are provided. Subsequently, these dominance rules are used in a dynamic programming algorithm, as well as a mixed integer linear programming formulation, with which this clustering problem can be solved to optimality. To conclude this section, both methods are illustrated by means of a numerical example.

2.4.1 Example

Consider the clustering problem with shared set-ups, as shown in Figure 2.3b. Then the costs $\lambda(U)$ for all possible clusters $U \subseteq \mathcal{J}$ are determined as follows:

$$\begin{gathered}
\lambda(\{1\}) = 5 \cdot (50 + 40 + 10) = 500, \\
\lambda(\{2\}) = 3 \cdot (50 + 40 + 20) = 330, \\
\lambda(\{3\}) = 2 \cdot (50 + 30) = 160, \\
\lambda(\{1, 2\}) = 5 \cdot (50 + 40 + 10 + 20) = 600, \\
\lambda(\{1, 3\}) = 5 \cdot (50 + 40 + 10 + 30) = 650, \\
\lambda(\{2, 3\}) = 3 \cdot (50 + 40 + 20 + 30) = 420, \\
\lambda(\{1, 2, 3\}) = 5 \cdot (50 + 40 + 10 + 20 + 30) = 750.
\end{gathered}$$

In a similar way, the costs $\Lambda(\Omega)$ for all possible clusterings Ω of $\mathcal{J}$ are given by:

$$\Lambda(\{\{1\},\{2\},\{3\}\}) = 500 + 330 + 160 = 990,$$
$$\Lambda(\{\{1\},\{2,3\}\}) = 500 + 420 = 920,$$
$$\Lambda(\{\{2\},\{1,3\}\}) = 330 + 650 = 980,$$
$$\Lambda(\{\{3\},\{1,2\}\}) = 160 + 600 = 760,$$
$$\Lambda(\{\{1,2,3\}\}) = 750.$$

Apparently, the optimal clustering is determined by $\Omega^* = \{\{1,2,3\}\}$, with corresponding (minimal) costs $\Lambda(\Omega^*) = 750$.

2.4.2 Problem reduction

As in the clustering problem with common set-ups, let us now derive some dominance rules, with which optimal clusterings can be characterized, and as a result of which the complexity of the clustering problem can be reduced significantly. As a starting point of our analysis, we denote with $s_{jk} > 0$ the shared set-up costs of maintenance jobs $j, k \in \mathcal{J}$:

$$s_{jk} = \sum_{i \in I_j \cap I_k} s_i.$$

Note that $s_{jk} \geq s_1 > 0$ for all $j, k \in \mathcal{J}$, since we assumed the existence of exactly one common set-up activity $i = 1$, with corresponding costs $s_1 > 0$. With this in mind, Theorem 1 can now be generalized as follows. But first, we need the following lemma.

Lemma 1 *Consider an arbitrary clustering Ω, and let $Q_j \in \Omega$ denote the cluster corresponding to maintenance job $j \in \mathcal{J}$. Furthermore, define $\delta_{jk} = (s_{jj} + c_j - s_{jk})/c_j$ for all $j, k \in \mathcal{J}$. Then $\delta_{kj}^{-1} \cdot f(Q_k) > f(Q_j) \geq f_k$ for some $j, k \in \mathcal{J}$ implies that Ω is not optimal.*

Proof. Suppose that $f(Q_j) \geq f_k$ for some $j, k \in \mathcal{J}$. Then removing job k from cluster Q_k results in a cost decrement of at least $\Delta^- = f(Q_k) \cdot c_k$. Similarly, moving job k to cluster Q_j results in a cost increment of at most $\Delta^+ = f(Q_j) \cdot (s_{kk} + c_k - s_{jk})$, since $f(Q_j) \geq f_k$ by assumption, and $j \in Q_j$ by definition. Since $\delta_{kj}^{-1} \cdot f(Q_k) > f(Q_j)$ is equivalent to $\Delta^- > \Delta^+$, this completes the proof. □

It is easily verified that $\delta_{jk} \geq 1$ for all $j, k \in \mathcal{J}$. Furthermore, $\delta_{jk} = \delta_{kj} = 1$ if $s_{jj} = s_{jk} = s_{kj} = s_{kk}$ for some $j, k \in \mathcal{J}$, i.e. if maintenance jobs j and k require the exact

same set-up activities. In general terms, Lemma 1 provides optimality conditions for the cluster frequencies $\{f(Q_j), f(Q_k)\}$ of each pair $\{j,k\}$ of maintenance jobs. It is now possible to generalize Theorem 1.

Theorem 2 *Consider an optimal clustering Ω^* and let $Q_j \in \Omega^*$ denote the cluster corresponding to maintenance job $j \in \mathcal{J}$. Then the following must be satisfied:*

$$
\begin{array}{ll}
(i) & \forall j,k \in \mathcal{J} : f(Q_j) = f(Q_k) \rightarrow Q_j = Q_k, \\
(ii) & \forall j,k \in \mathcal{J} : f_j \geq f_k \rightarrow f(Q_j) \geq \delta_{kj}^{-1} \cdot f(Q_k), \\
(iii) & \forall j \in \mathcal{J} : f(Q_j) \leq f_j \cdot (s_{jj} + c_j)/c_j, \\
(iv) & \forall j,k \in \mathcal{J} : f_j = f_k \rightarrow \delta_{jk}^{-1} \leq \frac{f(Q_k)}{f(Q_j)} \leq \delta_{kj}, \\
(v) & \forall j,k,l \in \mathcal{J} : f_j \geq f_l \geq f_k \ \wedge \ Q_j = Q_k \rightarrow \delta_{kl}^{-1} \leq \frac{f(Q_l)}{f(Q_j = Qk)} \leq \min\{\delta_{lj}, \delta_{lk}\}.
\end{array}
$$

Proof. If Ω^* violates (i), then $f(Q_j) = f(Q_k)$ and $Q_j \neq Q_k$ for some $j,k \in \mathcal{J}$. In that case, combination of clusters Q_j and Q_k results in a clustering Ω with $\Lambda(\Omega) \leq \Lambda(\Omega^*) - f(Q_j) \cdot s_{jk} < \Lambda(\Omega^*)$, since $s_{jk} > 0$ by assumption. If Ω^* violates (ii), then $f_j \geq f_k$ and $\delta_{kj}^{-1} \cdot f(Q_k) > f(Q_j)$ for some $j,k \in \mathcal{J}$. With $f(Q_j) \geq f_j \geq f_k$, this yields $\delta_{kj}^{-1} \cdot f(Q_k) > f(Q_j) \geq f_k$, and it follows from Lemma 1 that Ω^* is not optimal. If Ω^* violates (iii), then $f(Q_j) \cdot c_j > f_j \cdot (s_{jj} + c_j)$ for some $j \in \mathcal{J}$. In that case, removing job j from cluster Q_j and creating a new cluster $\{j\}$ results in a clustering Ω with $\Lambda(\Omega) \leq \Lambda(\Omega^*) - f(Q_j) \cdot c_j + f_j \cdot (s_{jj} + c_j) < \Lambda(\Omega^*)$. Once again, properties (iv) and (v) can be derived in an analogous way to properties (i) and (ii). $\square$

It can easily be verified that Theorem 2 is indeed a generalization of Theorem 1. In other words, the dominance rules for the clustering problem with common set-ups, can also be applied to each shared set-up activity in isolation. More specifically, if we denote with $j,k \in J_i^*$ two maintenance jobs that are attached to the same set-up activity $i \in \mathcal{I}$, then $f(Q_j) = f(Q_k) \rightarrow Q_j = Q_k$ and $f_j \geq f_k \rightarrow f(Q_j) \geq f(Q_k)$, but also $f_j = f_k \rightarrow Q_j = Q_k$. This is a potentially valuable insight, since these dominance rules may strongly reduce the complexity of the clustering problem with shared set-ups.

2.4.3 A dynamic programming algorithm

In this section, we will show that the clustering problem for maintenance jobs with shared set-ups can also be solved by means of a dynamic programming algorithm. According to dominance rule (i) of Theorem 2, different clusters must have different frequencies. In other words, the clustering problem can be interpreted as the assignment of set-up activities $i \in \mathcal{I}$ and maintenance jobs $j \in J_i$ to maintenance

frequencies $k \in \mathcal{K}_i$, in such a way that total (expected) costs are minimized. Recall that $\mathcal{K}_i \subseteq \mathcal{K}$ denotes the set of feasible maintenance frequencies for set-up activity $i \in \mathcal{I}$, whereas $\mathcal{K} = \{f_j \mid j \in \mathcal{J}\}$ denotes the set of different maintenance frequencies.

Our analysis now proceeds as follows. As a starting point, we denote with $g_i(K)$ the minimal costs for the subtree associated with set-up activity $i \in \mathcal{I}$, provided that set-up activities and maintenance jobs within this subtree can only be assigned to frequencies $k \in K \subseteq \mathcal{K}_i$. For each set-up activity $i \in \mathcal{I}$, and each possible state $K \subseteq \mathcal{K}_i$, we must now decide which frequencies $L \subseteq K$ to use. Based upon this decision, the (optimal) assignment of maintenance jobs $j \in J_i^*$ to maintenance frequencies $l \in L$, as well as the consequences for all lower-level set-up activities $i' \in S_i$, are immediately clear. More specifically, $g_i(K)$ can be determined recursively by means of the following dynamic programming equation:

$$g_i(K) = \min_{L \in \mathcal{L}_i(K)} \left\{ h_i(L) + \sum_{i' \in S_i} g_{i'}(L \cap \mathcal{K}_{i'}) \right\}.$$

Here, $h_i(L)$ denotes the (minimal) costs associated with assigning set-up activity $i \in \mathcal{I}$ and maintenance jobs $j \in J_i^*$ to maintenance frequencies $l \in L$. Obviously, $h_i(L)$ can be determined rather straightforwardly by means of the following expression:

$$h_i(L) = \sum_{l \in L} s_i \cdot l + \sum_{j \in J_i^*} c_j \cdot \min\{l \in L \mid l \geq f_j\}.$$

Moreover, $\mathcal{L}_i(K)$ denotes the set of feasible decisions $L \subseteq K$ that can be made for set-up activity $i \in \mathcal{I}$ in state $K \subseteq \mathcal{K}_i$:

$$\mathcal{L}_i(K) = \{L \subseteq K \mid \max\{l \in L\} \geq \max\{f_j \mid j \in J_i\}\}.$$

Obviously, the optimal clustering is now determined by calculation of $g_1(\mathcal{K})$, since $i = 1$ denotes the common i.e. highest-level set-up activity, and $\mathcal{K}_1 = \mathcal{K}$ by definition. In the worst case, this yields a dynamic programming algorithm with $2^{|\mathcal{K}|}$ states and decisions, where $|\mathcal{K}|$ denotes the number of different maintenance frequencies. On the other hand, the set of feasible states $K \subseteq \mathcal{K}_i$ and decisions $L \subseteq K$ for set-up activity $i \in \mathcal{I}$ can often be reduced significantly, by observing that the only relevant decisions $L \in \mathcal{L}_i(K)$ are those that satisfy the following condition:

$$L, L' \in \mathcal{L}_i(K) \Rightarrow h_i(L) < h_i(L') \text{ or } L \cap \mathcal{K}_{i'} \not\subseteq L' \cap \mathcal{K}_{i'} \text{ for some } i' \in S_i.$$

The underlying observation behind this reasoning is that a relevant decision $L \subseteq K$ should not be outperformed by another relevant decision $L' \subseteq K$ under all circumstances. At the very least, this means that $h_i(L) \geq h_i(L')$ for some $L \subset L' \subseteq K$

implies that decision $L \subseteq K$ can as well be neglected. Despite the potential reductions obtained this way, it is still immediately clear that this dynamic programming algorithm becomes inattractive, or even intractable, if the number of maintenance frequencies grows too large (e.g. $|\mathcal{K}| > 10$). In such cases, it seems worthwile to explore the possibilities for the use of other methods. To this end, we also developed a mixed integer linear programming formulation.

2.4.4 A mixed integer linear programming formulation

Let us now present a mixed integer linear programming formulation, which can be used to determine an optimal clustering for maintenance jobs with shared set-ups. To this end, the assignment of set-up activities $i \in \mathcal{I}$ to maintenance frequencies $k \in \mathcal{K}_i$ is comprised into variables $x_{ik} \in \{0,1\}$. In a similar way, the assignment of maintenance jobs $j \in \mathcal{J}$ to maintenance frequencies $k \in K_j$ is represented by variables $y_{jk} \in \{0,1\}$:

$$x_{ik} = \begin{cases} 1 & \text{if set-up } i \in \mathcal{I} \text{ is assigned to frequency } k \in \mathcal{K}_i, \\ 0 & \text{otherwise,} \end{cases}$$

$$y_{jk} = \begin{cases} 1 & \text{if job } j \in \mathcal{J} \text{ is assigned to frequency } k \in K_j, \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $K_j = \{k \in \mathcal{K} \mid k \geq f_j\}$ reflects the set of feasible maintenance frequencies for maintenance job $j \in \mathcal{J}$. With $a_{ik} = k \cdot s_i > 0$, we denote the average costs per unit of time associated with the assignment of set-up activity $i \in \mathcal{I}$ to maintenance frequency $k \in \mathcal{K}_i$. Similarly, $b_{jk} = k \cdot c_j > 0$ denotes the average costs per unit of time associated with the assignment of component $j \in \mathcal{J}$ to maintenance frequency $k \in K_j$. Since $x_{ik} = \max\{y_{jk} \mid j \in J_i\}$ by definition, our problem can now be formulated as a mixed integer linear program, where our objective is to minimize the (expected) total costs per unit of time:

$$\text{Minimize } z = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}_i} a_{ik} \cdot x_{ik} + \sum_{j \in \mathcal{J}} \sum_{k \in K_j} b_{jk} \cdot y_{jk}$$

Subject to:

$$\begin{array}{lrcll}
\text{(a)} & & x_{ik} & \geq & x_{i'k} & \forall i \in \mathcal{I}, i' \in S_i, k \in \mathcal{K}_{i'} \\
\text{(b)} & & x_{ik} & \geq & y_{jk} & \forall i \in \mathcal{I}, j \in J_i^*, k \in K_j \\
\text{(c)} & \sum_{k \in K_j} & y_{jk} & = & 1 & \forall j \in \mathcal{J} \\
\text{(d)} & & x_{ik} & \geq & 0 & \forall i \in \mathcal{I}, k \in \mathcal{K}_i \\
\text{(e)} & & y_{jk} & \in & \{0,1\} & \forall j \in \mathcal{J}, k \in K_j
\end{array}$$

Here, restriction (a) prescribes that the assignment of a set-up activity $i \in \mathcal{I}$ to a maintenance frequency $k \in \mathcal{K}_i$ requires the parental set-up activities to be carried out with the same maintenance frequency too. In a similar way, restriction (b) prescribes that the assigment of maintenance job $j \in J_i^*$ to maintenance frequency $k \in K_j$ requires the corresponding set-up activity $i \in \mathcal{I}$ to be carried out with the same maintenance frequency too. Moreover, restrictions (c) and (e) guarantee that exactly one frequency $k \in K_j$ is assigned to each maintenance job $j \in \mathcal{J}$. As a consequence, restriction (d) is sufficient to ensure that $x_{ik} \in \{0, 1\}$ for all $i \in \mathcal{I}$ and $k \in \mathcal{K}_i$.

2.4.5 Example (continued)

Let us now illustrate the above-mentioned methods by reconsidering the example of Figure 2.3b. As a starting point, the **dynamic programming algorithm** determines the set of relevant decisions $\mathcal{L}_1(\mathcal{K}_1)$ for set-up activity $i = 1$. The only feasible decisions are given by $\{5\}$, $\{5, 3\}$, $\{5, 2\}$ and $\{5, 3, 2\}$, with corresponding costs:

$$\begin{aligned} h_1(\{5\}) &= 5 \cdot 50 + 5 \cdot 30 = 400, \\ h_1(\{5, 3\}) &= (5 + 3) \cdot 50 + 3 \cdot 30 = 490, \\ h_1(\{5, 2\}) &= (5 + 2) \cdot 50 + 2 \cdot 30 = 410, \\ h_1(\{5, 3, 2\}) &= (5 + 3 + 2) \cdot 50 + 2 \cdot 30 = 560. \end{aligned}$$

Since $\mathcal{K}_2 = \{5, 3\}$, it is easily verified that decisions $\{5, 2\}$ and $\{5, 3, 2\}$ can as well be neglected. After all, $h_1(\{5\}) < h_1(\{5, 2\})$ and $h_1(\{5, 3\}) < h_1(\{5, 3, 2\})$, whereas $\{5\} \cap \{5, 3\} = \{5, 2\} \cap \{5, 3\} = \{5\}$ and $\{5, 3\} \cap \{5, 3\} = \{5, 3, 2\} \cap \{5, 3\} = \{5, 3\}$. Apparently, the only relevant decisions for set-up activity $i = 1$ are $\{5\}$ and $\{5, 3\}$, i.e. $\mathcal{L}_1(\mathcal{K}_1) = \{\{5\}, \{5, 3\}\}$. Consequently, the only relevant states for set-up activity $i = 2$ are $K = \{5\} \cap \{5, 3\} = \{5\}$ and $K = \{5, 3\} \cap \{5, 3\} = \{5, 3\}$. With this in mind, the dynamic programming algorithm results in:

$$\begin{aligned} g_2(\{5\}) &= h_2(\{5\}) \\ &= 5 \cdot 40 + 5 \cdot (10 + 20) = 350, \\ g_2(\{5, 3\}) &= h_2(\{5, 3\}) \\ &= (5 + 3) \cdot 40 + 5 \cdot 10 + 3 \cdot 20 = 430, \\ g_1(\{5, 3, 2\}) &= \min\{h_1(\{5\}) + g_2(\{5\}), h_2(\{5, 3\}) + g_2(\{5, 3\})\} \\ &= \min\{400 + 350, 490 + 430\} \\ &= \min\{750, 920\} = 750. \end{aligned}$$

In a similar way, the **mixed integer linear program** starts with $\mathcal{K}_1 = \{5, 3, 2\}$, $\mathcal{K}_2 = \{5, 3\}$, $K_1 = \{5\}$, $K_2 = \{5, 3\}$ and $K_3 = \{5, 3, 2\}$, and consequently defines

11 variables x_{15}, x_{13}, x_{12}, x_{25}, x_{23}, y_{15}, y_{25}, y_{23}, y_{35}, y_{33} and y_{32}. While solving this problem with a standard ILP solver, the following (optimal) solution was found: $x_{15} = x_{25} = y_{15} = y_{25} = y_{35} = 1$ and $x_{13} = x_{12} = x_{23} = y_{23} = y_{33} = y_{32} = 0$, with corresponding costs $z = 750$. Surprisingly enough, this solution was also found if the integrality constraints (e) are relaxed. In section 2.6, this phenomenon will be studied in more detail. But first, we will present some efficient heuristics for the clustering problem with shared set-ups.

2.5 Efficient heuristics for the clustering problem

In this section, we will present two efficient heuristics for the clustering problem with shared set-ups, which are both capable of dealing with a large number of set-up activities, maintenance jobs and limitative frequencies (e.g. $|\mathcal{I}| = 100$, $|\mathcal{J}| = 1000$, $|\mathcal{K}| = 50$). To achieve this, these heuristics explore the nice structural properties of the clustering problem with common set-ups, in order to arrive at an optimal or near-optimal solution for the clustering problem with shared set-ups.

2.5.1 A top-down heuristic

The top-down heuristic (TDH) is based on the assumption that dominance rules (i) and (ii) for the clustering problem with common set-ups, as derived in Theorem 1, are also applicable to the clustering problem with shared set-ups. If we denote with $Q_j \subseteq \mathcal{J}$ the cluster corresponding to maintenance job $j \in \mathcal{J}$, these dominance rules impy that $f(Q_j) = f(Q_k) \rightarrow Q_j = Q_k$ and $f_j \geq f_k \rightarrow f(Q_j) \geq f(Q_k)$ for all $j, k \in \mathcal{J}$. Once again, this means that maintenance jobs j and k with the same frequency $f_j = f_k$ can as well be replaced by a single maintenance job $l = \{j\} \cup \{k\}$ with frequency $f_l = f_j = f_k$. By doing so, the collection of maintenance jobs is eventually subdivided into exactly $|\mathcal{K}|$ predefined clusters U_k $(1 \leq k \leq |\mathcal{K}|)$, each of which corresponds to a different maintenance frequency $k \in \mathcal{K}$. Subsequently, these clusters are ordered in such a way that $f(U_1) > ... > f(U_{|\mathcal{K}|})$, and the optimal clustering is determined recursively by means of the following dynamic programming equation:

$$g(k) = \min_{j \leq k} \{g(j-1) + \lambda(U_j \cup ... \cup U_k)\}.$$

Here, $g(k)$ denotes the minimal costs for clustering $U_1 ... U_k$, and we define $g(0) = 0$ for notational convience. Recall that $\lambda(U) = f(U) \cdot \{s(U) + c(U)\}$ denotes the (expected) costs associated with a cluster $U \subseteq \mathcal{J}$.

2.5.2 A bottom-up heuristic

In the bottom-up heuristic (BUH), the clustering problem with m shared set-ups is decomposed into m consecutive clustering problems with common set-ups. To be specific, the heuristic starts with the lowest-level set-up activities $\mathcal{L} = \{i \in \mathcal{I} \mid S_i = \emptyset\}$, and determines an optimal clustering for each lowest-level set-up activity $i \in \mathcal{L}$ and corresponding maintenance jobs $j \in J_i^*$ in isolation. Subsequently, the clusters obtained this way are used as individual composite maintenance jobs at all higher levels, and this recursive procedure is repeated by proceeding upwards to the root of the maintenance tree ($i = 1$). For each set-up activity $i \in \mathcal{I}$ in isolation, we assume that dominance rules (i) and (ii) for the clustering problem with common set-ups, as derived in Theorem 1, can be applied.

As an illustrative example, consider an arbitrary set-up activity $i \in \mathcal{I}$, and assume that predefined clusters U_1 t/m U_p with $U_1 \cup ... \cup U_p = J_i$ have been formed by the lower-level set-up activities $i' \in S_i$, and by the individual maintenance jobs $j \in J_i^*$. As before, clusters U_j and U_k with identical frequencies $f(U_j) = f(U_k)$ are now replaced by a single cluster $U_j \cup U_k$ with frequency $f(U_j \cup U_k) = f(U_j) = f(U_k)$. Subsequently, the remaining $q \leq p$ clusters U_k $(1 \leq k \leq q)$ are ordered in such a way that $f(U_1) > ... > f(U_q)$. Once again, the optimal clustering for the subtree associated with set-up activity $i \in \mathcal{I}$, is now determined recursively by means of the following dynamic programming equation:

$$g(k) = \min_{j \leq k} \{g(j-1) + \lambda_i(U_j \cup ... \cup U_k)\}.$$

Here, $g(k)$ denotes the minimal costs for clustering $U_1 ... U_k$, and we define $g(0) = 0$ for notational convience. Moreover, $\lambda_i(U) \leq \lambda(U)$ reflects the costs associated with a cluster $U \subseteq \mathcal{J}$, exclusive of the costs of all higher level set-up activities $i' < i$ (e.g. $\lambda_2(\{1,2\}) = 5 \cdot (40 + 10 + 20) = 350$ in Figure 2.3b). For notational convenience, we will denote with Ω_i the optimal clustering of the subtree associated with set-up activity $i \in \mathcal{I}$, and with $\Lambda_i(\Omega_i)$ the corresponding (minimal) costs in the sequel.

2.5.3 Example (continued)

As a starting point, the **top-down heuristic** starts with predetermined clusters $U_1 = \{1\}$, $U_2 = \{2\}$, and $U_3 = \{3\}$, with cluster frequencies $f(U_1) = 5$, $f(U_2) = 3$ and $f(U_3) = 2$. Subsequently, the dynamic programming algorithm arrives at $\Omega = \{\{1,2,3\}\}$, with corresponding costs $\Lambda(\Omega) = 5 \cdot (50 + 40 + 10 + 20 + 30) = 750$. In a similar way, the **bottom-up heuristic** starts with the lowest level set-up activity

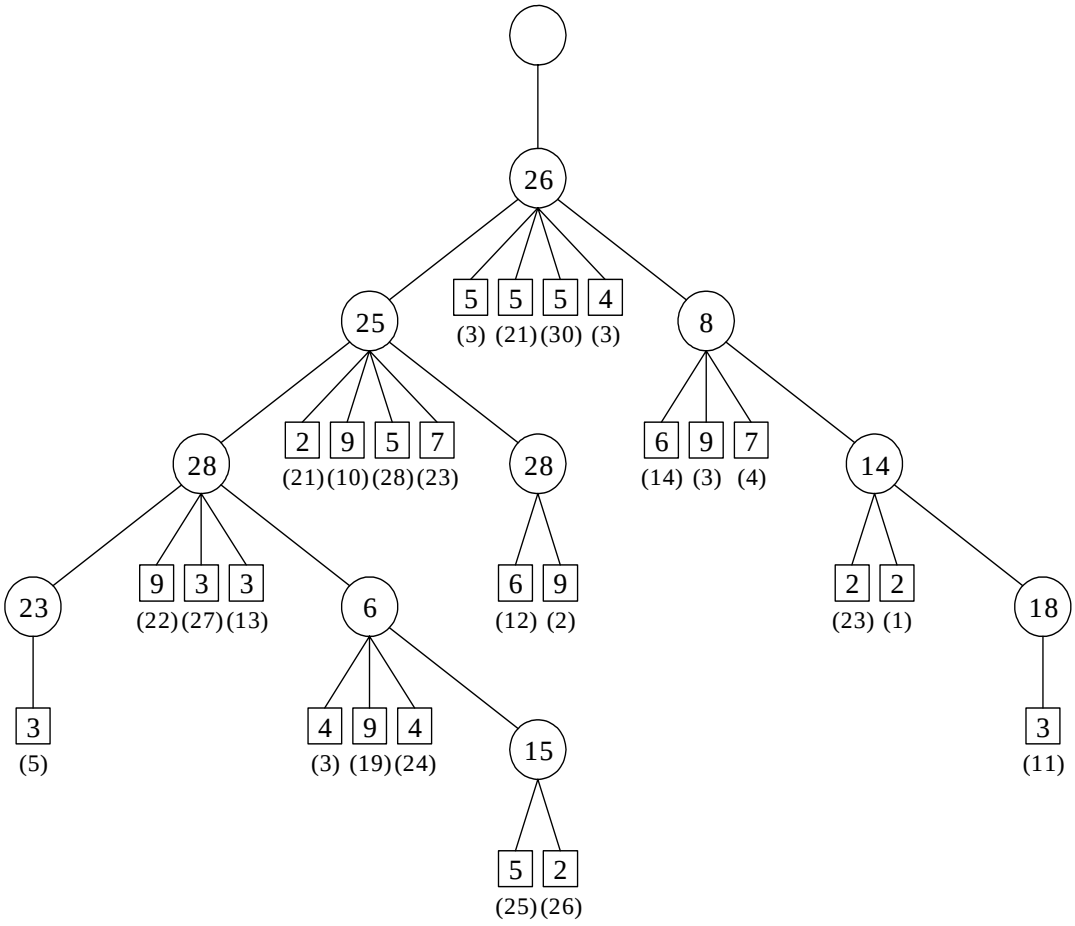


Figure 2.4: Example of a test problem with 10 set-up activities and 25 maintenance jobs: set-up and maintenance costs are shown at the corresponding nodes, limitative frequencies in brackets below.

$i = 2$, with predetermined clusters $U_1 = \{1\}$ and $U_2 = \{2\}$, and cluster frequencies $f(U_1) = 5$ and $f(U_2) = 3$. Subsequently, the dynamic programming algorithm arrives at $\Omega_2 = \{\{1, 2\}\}$, with corresponding costs $\Lambda_2(\Omega_2) = 5 \cdot (40 + 10 + 20) = 350$. As a next step, set-up activity $i = 1$ is considered with predetermined clusters $U_1 = \{1, 2\}$ and $U_2 = \{3\}$, and cluster frequencies $f(U_1) = 5$ and $f(U_2) = 2$. In a similar way, the dynamic programming arrives at $\Omega_1 = \{\{1, 2, 3\}\}$, with corresponding costs $\Lambda_1(\Omega_1) = 5 \cdot (50 + 40 + 10 + 20 + 30) = 750$.

2.6 Computational results

Let us now present the results of a series of numerical experiments that were carried out to investigate the performance of both heuristics, in relation to the optimal so-

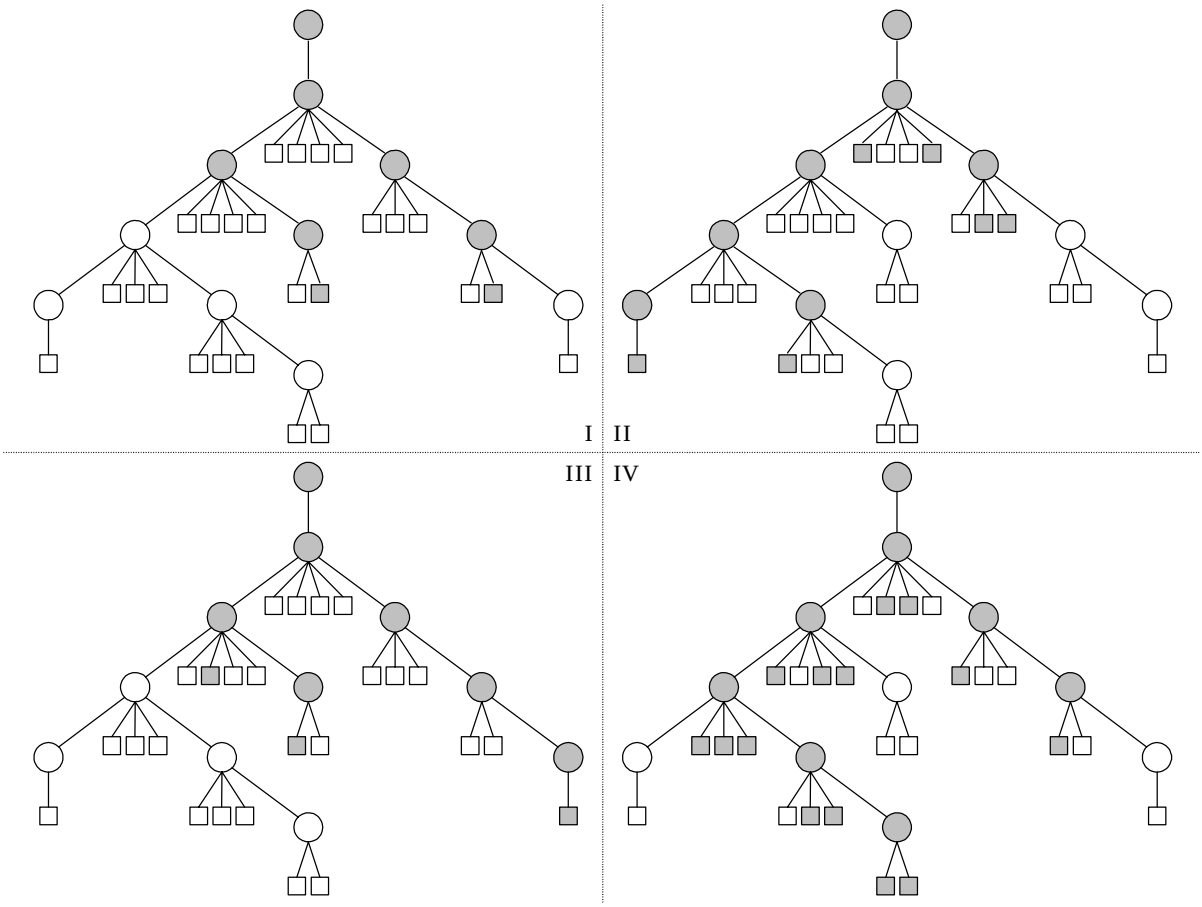


Figure 2.5: Optimal clustering for the test problem of Figure 2.4.

lution provided by the mixed integer linear programming formulation. To this end, we created a series of 1000 instances for 24 types of test problems, each of which corresponds to a fixed number of set-up activities $m \in \{5, 10\}$, a fixed number of maintenance jobs $n \in \{25, 50\}$, a maximal set-up cost $\bar{s} \in \{10, 20, 30\}$, a maximal maintenance cost $\bar{c} \in \{10, 20, 30\}$, and a maximal limitative frequency $\bar{f} \in \{15, 30\}$. In each test problem, the parent of set-up activity $i \in \{1, ..., m\}$ was drawn at random from the set $\{1, ..., i-1\}$. In a similar way, the parental set-up activity of maintenance job $j \in \{1, ..., n\}$ was chosen randomly from the set $\{1, ..., m\}$. Finally, the parameters s_i, c_j and f_j were drawn at random from the sets $\{1, ..., \bar{s}\}$, $\{1, ..., \bar{c}\}$ and $\{1, ..., \bar{f}\}$ respectively (see Figure 2.4 for an example).

For each test problem obtained this way, we determined the optimal solution with the use of the mixed integer linear programming formulation of section 2.4. Moreover, we administrated the fraction of times that its LP relaxation arrived at an integer optimal solution, and the number of clusters in the optimal solution as well (see Figure

2.5 for an example). The performance of both heuristics was determined in terms of the fraction of times that they were optimal, and the average resp. maximal deviation from the optimal solution as well. In addition, a pairwise comparison between both heuristics was made. The results are depicted in Tables 2.3, 2.4 and 2.5 respectively.

From Table 2.3, it can be observed that the vast majority of clustering problems could be solved to optimality by means of a linear programming formulation. In all other cases, the deviations were relatively small, with a maximum of only 0.38% over all 24.000 test problems. Moreover, an integer optimal solution could usually be found within a few iterations of the consecutive branch and bound algorithm. Finally, and in compliance with our general expectations, the number of clusters in the optimal solution increases with the number of maintenance jobs and limitative frequencies, but decreases with the ratio of set-up to maintenance costs.

In a similar way, the results in Table 2.4 indicate that the average performance of both heuristics decreases with the number of set-up activities, the number of maintenance jobs, the number of limitative frequencies, and the ratio of set-up to maintenance costs. At the same time, however, the performance of both heuristics in terms of the fraction of times that they were optimal increases with the ratio of set-up to maintenance costs. In our opinion, this counter-intuitive behavior can be explained by observing that an optimal clustering consisting of only a few clusters (due to high set-up costs) is more likely to be generated by the heuristics, in comparison with an optimal clustering consisting of multiple clusters (due to low set-up costs).

Summarizing, we conclude that the bottom-up heuristic clearly outperforms the top-down heuristic, at least from an overall perspective. This does not necessarily mean, however, that the bottom-up heuristic outperforms the top-down heuristic for each individual test problem. The results in Table 2.5 provide a comparitive study into the relative performance of both heuristics. Apparently, the top-down heuristic outperforms the bottom-up heuristic in a significant part of all test problems, although its relative performance reduces strongly as the number of set-up activities grows. This is intuitively clear, since the bottom-up heuristic more explicitly takes into account the complexity of the maintenance tree under consideration. Nevertheless, we believe that both heuristics are of practical value, because each of them generates near-optimal solutions within negligible computation times.

2.7 Concluding remarks

In this chapter, we showed that the clustering problem for frequency-constrained maintenance jobs with common set-ups can be solved in polynomial time by means

Table 2.3: Outcomes of the (mixed integer) linear programming formulation for 24×1000 randomly generated test problems with m set-up activities and n maintenance jobs, where $s_i \in \{1, ..., \bar{s}\}$, $c_j \in \{1, ..., \bar{c}\}$ and $f_j \in \{1, ..., \bar{f}\}$.

					% deviation (LP)		# clusters (MILP)		
m	n	$\overline{s}$	$\overline{c}$	$\overline{f}$	optimal	maximal	minimal	average	maximal
5	25	10	30	15	99.3	0.22	3	6.73	13
5	25	20	20	15	99.0	0.41	2	4.74	11
5	25	30	10	15	99.6	0.14	1	3.17	9
5	25	10	30	30	98.6	0.20	3	7.17	16
5	25	20	20	30	99.2	0.26	2	4.84	12
5	25	30	10	30	99.3	0.26	1	3.32	11
5	50	10	30	15	98.0	0.15	5	9.15	15
5	50	20	20	15	96.7	0.37	3	6.48	13
5	50	30	10	15	98.4	0.11	2	4.35	11
5	50	10	30	30	94.6	0.21	5	10.47	21
5	50	20	20	30	95.5	0.24	3	6.96	17
5	50	30	10	30	97.3	0.23	2	4.58	15
10	25	10	30	15	99.0	0.26	3	6.58	13
10	25	20	20	15	98.9	0.18	1	4.79	12
10	25	30	10	15	99.6	0.11	1	3.19	8
10	25	10	30	30	98.6	0.21	2	6.95	17
10	25	20	20	30	99.1	0.36	1	4.85	12
10	25	30	10	30	99.8	0.10	1	3.37	9
10	50	10	30	15	97.5	0.15	4	8.86	15
10	50	20	20	15	96.0	0.23	2	6.25	14
10	50	30	10	15	97.7	0.38	1	4.27	11
10	50	10	30	30	95.6	0.17	4	10.27	20
10	50	20	20	30	95.2	0.30	3	6.94	16
10	50	30	10	30	98.2	0.16	1	4.54	12

Table 2.4: Performance of the top-down (TDH) and bottom-up (BUH) heuristic, for 24×1000 randomly generated test problems with m set-up activities and n maintenance jobs, where $s_i \in \{1, ..., \bar{s}\}$, $c_j \in \{1, ..., \bar{c}\}$ and $f_j \in \{1, ..., \bar{f}\}$.

					% deviation (TDH)			% deviation (BUH)		
m	n	$\overline{s}$	$\overline{c}$	$\overline{f}$	optimal	average	maximal	optimal	average	maximal
5	25	10	30	15	16.1	1.0	7.1	51.7	0.2	3.4
5	25	20	20	15	18.2	1.7	11.7	46.5	0.4	4.7
5	25	30	10	15	25.5	2.1	16.3	56.0	0.4	4.9
5	25	10	30	30	17.9	0.9	6.3	46.9	0.2	2.8
5	25	20	20	30	16.0	1.7	11.8	44.0	0.4	4.4
5	25	30	10	30	25.2	2.0	14.0	47.6	0.4	3.8
5	50	10	30	15	4.4	1.0	4.3	30.1	0.3	2.6
5	50	20	20	15	6.1	1.7	8.8	25.4	0.5	3.9
5	50	30	10	15	15.1	2.1	12.5	28.1	0.7	5.7
5	50	10	30	30	4.0	1.0	4.5	19.4	0.3	2.3
5	50	20	20	30	7.7	1.6	8.3	19.5	0.5	3.7
5	50	30	10	30	11.4	2.0	12.6	24.6	0.6	4.7
10	25	10	30	15	6.8	1.5	9.0	37.0	0.3	4.4
10	25	20	20	15	7.3	2.7	14.9	39.1	0.5	6.1
10	25	30	10	15	13.8	3.2	17.9	50.3	0.4	6.1
10	25	10	30	30	7.3	1.4	8.0	35.0	0.3	3.6
10	25	20	20	30	7.0	2.6	13.6	35.5	0.5	7.1
10	25	30	10	30	13.2	3.2	19.4	44.3	0.5	5.1
10	50	10	30	15	1.0	1.7	7.3	19.9	0.4	2.1
10	50	20	20	15	1.4	2.7	10.5	14.1	0.7	2.7
10	50	30	10	15	5.0	3.2	14.3	24.0	0.6	3.2
10	50	10	30	30	0.2	1.7	5.6	10.6	0.4	1.7
10	50	20	20	30	1.7	2.8	12.2	12.7	0.6	2.8
10	50	30	10	30	3.8	3.1	15.3	19.0	0.6	3.1

Table 2.5: Comparison of the top-down (TDH) and bottom-up (BUH) heuristic, for 24×1000 randomly generated test problems with m set-up activities and n maintenance jobs, where $s_i \in \{1, ..., \bar{s}\}$, $c_j \in \{1, ..., \bar{c}\}$ and $f_j \in \{1, ..., \bar{f}\}$.

m	n	$\overline{s}$	$\overline{c}$	$\overline{f}$	% TDH < BUH	% TDH = BUH	% TDH > BUH
5	25	10	30	15	15.6	12.5	71.9
5	25	20	20	15	17.9	12.8	69.3
5	25	30	10	15	14.6	21.3	64.1
5	25	10	30	30	17.2	15.3	67.5
5	25	20	20	30	17.5	13.2	69.3
5	25	30	10	30	15.6	23.0	61.4
5	50	10	30	15	18.2	1.7	80.1
5	50	20	20	15	21.3	1.9	76.8
5	50	30	10	15	27.4	5.2	67.4
5	50	10	30	30	20.1	1.2	78.7
5	50	20	20	30	25.0	1.9	73.1
5	50	30	10	30	24.7	5.3	70.0
10	25	10	30	15	11.2	7.8	81.0
10	25	20	20	15	11.3	7.1	81.6
10	25	30	10	15	7.5	16.1	76.4
10	25	10	30	30	11.3	8.9	79.8
10	25	20	20	30	11.7	8.1	80.2
10	25	30	10	30	8.0	15.9	76.1
10	50	10	30	15	11.5	0.6	87.9
10	50	20	20	15	13.9	0.8	85.3
10	50	30	10	15	12.7	3.7	83.6
10	50	10	30	30	10.6	0.1	89.3
10	50	20	20	30	11.6	0.5	87.9
10	50	30	10	30	13.1	3.4	83.5

of an efficient dynamic programming algorithm. In addition, we developed a dynamic programming algorithm, as well as a mixed integer linear programming formulation, which can be used to determine an optimal clustering for frequency-constrained maintenance jobs with shared set-ups. Finally, two efficient heuristics were developed which generate near-optimal solutions in negligible computation times. Summarizing, we believe that our methods can be applied in many practical situations, and are a step forward into effective and efficient strategic maintenance planning. Nevertheless, a number of possible generalizations of, and extensions to our modelling framework are still under consideration.

As a starting point, the positive effect of preventive maintenance clustering in terms of a reduction in equipment failures and corresponding corrective maintenance costs, was not considered in our analysis. In this respect, it seems worthwile to consider frequency-dependent costs instead of frequency constraints, as a result of which clustering might become even more profitable in comparison with our approach. Another possibility is to allow parallel execution of maintenance jobs within a cluster, and/or simultaneous execution of clusters within a clustering. In this chapter, we assumed that no cost reductions could be obtained by carrying out maintenance jobs in parallel, or maintenance packages simultaneously. Of course, other assumptions would lead to other interesting versions of the clustering problem. We will come back to that in the following chapter.

Finally, it might also be worthwile to incorporate the possibilities for workload balancing in our modelling framework. In view of cost efficiency, it might be optimal to carry out all maintenance jobs simultaneously, whereas this solution might be less attractive from an administrative point of view. To avoid this, an additional restriction could be build in concerning the total costs and/or times of each individual cluster. Moreover, sensitivity analysis could be performed in order to determine the extra costs that would be incurred due to this additional constraint on the clustering problem. These suggestions, however, are left for future research.

Chapter 3

Coordinated planning of preventive maintenance jobs with shared set-ups and frequency-dependent costs

In this chapter, we consider an indirect clustering problem for preventive maintenance jobs with shared set-ups. More specifically, we consider a multi-setup multi-component production system, in which each component is maintained preventively at integer multiples of a certain basis interval. Once again, creating an occasion for preventive maintenance on one of these components requires a collection of preparatory set-up activities to be carried out in advance, with corresponding set-up costs. The main difference with the previous chapter, however, is that corrective maintenance costs are also incorporated. This leads to frequency-dependent maintenance costs rather than frequency constraints. A general mathematical framework is presented which allows for a large class of preventive maintenance strategies for each component. Our approach generalizes previous work in the sense that it provides efficient MILP and DP formulations, in which considerably more degrees of freedom are taken into account. Computational results indicate that near-optimal solutions are obtained within reasonable computation times.

3.1 Introduction

As in the previous chapter, we consider a production system consisting of multiple set-up activities and multiple components, which are organized in a tree-like structure. Creating an occasion for preventive maintenance on one of these components, requires a collection of preparatory set-up activities to be carried out in advance, with

corresponding set-up costs. If components are maintained simultaneously, the corresponding set-up activities can be combined. In this respect, there is a perspective of significant savings if some kind of coordination is built in concerning the starting time of preventive maintenance on different components. In the previous chapter, this was incorporated by means of a so-called direct clustering problem, in which the collection of maintenance jobs was subdivided into several maintenance packages.

In this chapter, clusters of maintenance jobs are formed indirectly whenever possible. More specifically, we consider the case where each component is maintained preventively at integer multiples of a certain basis interval, which is the same for all components. In general, this yields an optimization problem in $n+1$ variables, where n denotes the number of components. The main difference between the previous chapter and our approach here, is that corrective maintenance costs are also incorporated. This leaves us with frequency-dependent maintenance costs rather than frequency constraints. The underlying idea behind this approach originates from Gertsbakh (1977), who considered a somewhat similar but less powerful modelling framework with multiple set-up activities, in which components can only be attached to lowest-level set-up activities (see Figure 3.1). Since then, applications of this type of set-up structures into maintenance modelling have been scarce, e.g. see Sculli and Suraweera (1979) and Van Dijkhuizen and Van Harten (1997b), possibly because of both practical and theoretical complications.

At the same time, the single set-up version of this problem has gained considerably more attention in existing literature, since it is closely related to the well-known **joint replenishment** problem from multi-item inventory theory, e.g. see Bomberger (1966), Goyal (1973), Goyal (1974), Silver (1976), Kaspi and Rosenblatt (1983), Hariga (1994) and Ben-Daya and Hariga (1995). The reader is referred to Goyal and Satir (1989) and Wildeman, Frenk, and Dekker (1997) for extensive and up-to-date literature reviews on the joint replenishment problem. Here, we restrict ourselves to its applications into maintenance modelling. Pioneering work on this subject was carried out by Goyal and Kusy (1985) and Goyal and Gunasekaran (1992), who presented a number of iterative heuristics in case the deterioration cost functions are of a very specific form. Just recently, Wildeman (1996) presented a mathematical framework which allows for a much larger class of preventive maintenance models, and solved this problem to optimality.

In this chapter, we will present a more general and powerful modelling framework, which allows for multiple set-up activities, multiple components, and a variety of maintenance strategies for each component. Our approach generalizes previous work in the sense that it provides efficient dynamic programming and mixed integer linear

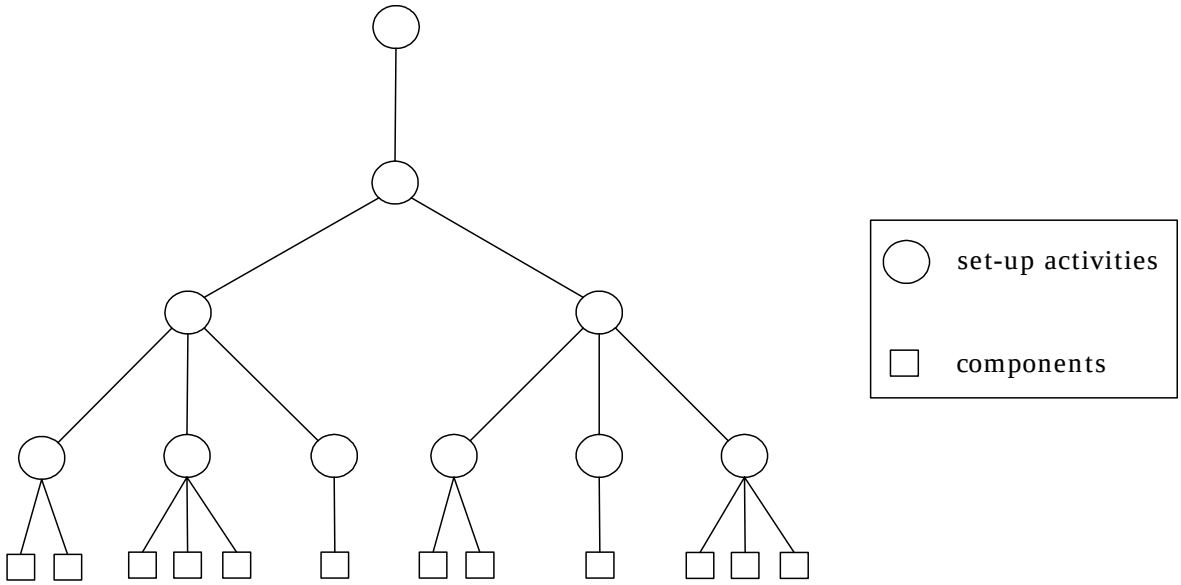


Figure 3.1: Example of the modelling framework considered by Gertsbakh (1977), in which components can only be attached to lowest-level set-up activities.

programming formulations, in which considerably more degrees of freedom are taken into account than in existing literature. Moreover, it contains some interesting new elements, which are typical for our setting of multiple set-up activities. But first, let us discuss our general approach in some more detail.

3.2 General approach

Consider a production system consisting of m set-up activities, denoted $\mathcal{I} = \{1, ..., m\}$, and n components, denoted $\mathcal{J} = \{1, ..., n\}$, which are organized in a tree-like structure. Creating an opportunity for preventive maintenance on component $j \in \mathcal{J}$ requires a collection $I_j \subseteq \mathcal{I}$ of preparatory set-up activities to be carried out in advance, with corresponding set-up costs $\sum_{i \in I_j} s_i$. Here, $s_i > 0$ denotes the individual cost of set-up activity $i \in \mathcal{I}$. If components are maintained simultaneously, the corresponding set-up activities can be combined. More specifically, preventive maintenance on a subset of components $U \subseteq \mathcal{J}$ involves a set-up cost $s(U)$, which depends completely on the set of required set-up activities. As in the previous chapter, this yields the following expression for $s(U)$:

$$s(U) = \sum_{i \in \bigcup\limits_{j \in U} I_j} s_i.$$

Within this setting, we consider preventive maintenance activities of the **block type**. To be specific, each component is maintained preventively at fixed intervals (e.g. daily, weekly, monthly, yearly), whereas intermediate corrective maintenance activities are carried out upon failure. With $\Phi_j(x)$, we denote the expected maintenance costs (exclusive of preventive set-up costs) of component $j \in \mathcal{J}$ per unit of time, if maintained preventively every $x > 0$ time units (or any other measurable quantity e.g. running hours). For notational convenience, and without loss of generality, we restrict ourselves to cost functions of the following type:

$$\Phi_j(x) = \frac{c_j + M_j(x)}{x}.$$

Here, $c_j > 0$ reflects the expected cost of preventive maintenance on component $j \in \mathcal{J}$. Moreover, $M_j(x)$ denotes the expected cumulative deterioration costs (due to failures, repairs, etc.), x time units after its last preventive maintenance. By doing this, a variety of maintenance models can be incorporated, allowing different models for each component. The reader is referred to Dekker (1995) for an extensive list of block-type models. Here, we only mention some important ones:

- in the **standard block replacement** model (Barlow and Proschan 1965), a component is replaced correctively upon failure, and preventively at fixed intervals of length $x > 0$;
- in the **modified block replacement** model (Berg and Epstein 1976), a component is replaced correctively upon failure, and preventively at fixed intervals of length $x > 0$, but only if its age exceeds a certain treshold value $y < x$;
- in the **minimal repair** model (Barlow and Hunter 1960), a component is replaced preventively at fixed intervals of length $x > 0$, with intermediate failure repairs occuring whenever necessary, restoring the component into a state as good as before failure;
- in the **standard inspection** model (Barlow, Hunter, and Proschan 1963), a component is inspected preventively at fixed intervals of length $x > 0$, followed by a corrective replacement if it turns out to have failed upon inspection;
- in the **delay time** model (Christer and Waller 1984), a component passes through a visible defective state before it actually fails somewhat later; in line with this, the component is replaced correctively upon failure, and inspected preventively at fixed intervals of length $x > 0$, followed by a preventive replacement if it turns out to be defective at the time of inspection.

In our modelling framework, we assume that the initiation of preventive maintenance on different components is based on the same system parameter x, e.g. calendar or operating time, and that the deterioration cost functions $M_j(x)$ are all available. In the modified block replacement model, this means that the optimal value of y can be determined once the value of x is given. Moreover, we assume that the individual cumulative deterioration costs $M_j(.)$ for each component are totally independent of all other components. In other words, the possibilities for dynamic and/or opportunistic grouping, i.e. the combination of preventive with corrective maintenance activities in an operational planning phase, are not accounted for in our approach here. Of course, other assumptions would lead to other interesting versions of the problem under consideration, but are left for future research.

3.2.1 Further notation and assumptions

Throughout this paper, and analogous to the previous chapter, we will use the following notation and assumptions to characterize the mutual relationships between set-up activities and components. First of all, we denote with $J_i = \{j \in \mathcal{J} \mid i \in I_j\}$ and $J_i^* \subseteq J_i$ the set of components that require resp. are attached to set-up activity $i \in \mathcal{I}$. In a similar way, we let $S_i \subset \mathcal{I}$ and $S_i^* \subset \mathcal{I}$ denote the set-up activities $i' \in \mathcal{I}$ that require resp. are attached to set-up activity $i \in \mathcal{I}$. As an illustrative example, consider a production system consisting of $m = 3$ set-up activities and $n = 3$ components, and assume that component j requires all set-up activities $i \leq j$ to be carried out in advance. Then it is easily verified that:

$$\begin{gathered} I_1 = \{1\},\ I_2 = \{1,2\},\ I_3 = \{1,2,3\}, \\ J_1 = \{1,2,3\},\ J_2 = \{2,3\},\ J_3 = \{3\}, \\ J_1^* = \{1\},\ J_2^* = \{2\},\ J_3^* = \{3\}, \\ S_1 = \{2,3\},\ S_2 = \{3\},\ S_3 = \varnothing, \\ S_1^* = \{2\},\ S_2^* = \{3\},\ S_3^* = \varnothing. \end{gathered}$$

For notational convenience, and without loss of generality, we assume that $i' > i$ for all $i \in \mathcal{I}$ and $i' \in S_i$ in the sequel. Moreover, and in line with the above, we assume that $I_1 \cap ... \cap I_n = \{1\}$, i.e. there exists exactly one common set-up activity $i = 1$. If $I_1 \cap ... \cap I_n = \emptyset$, the problem can be decomposed into two or more (smaller) subproblems, which can be treated and solved separately. If $|I_1 \cap ... \cap I_n| = k > 1$, the common set-up activities $i \in \{1, ..., k\}$ can as well be replaced by a single common set-up activity, with corresponding set-up costs $s_1 + ... + s_k > 0$. For similar reasons, we can assume that $J_i \neq \varnothing$ for all $i \in \mathcal{I}$, since obviously set-up activities with no components can as well be neglected without affecting the problem.

Table 3.1: Example of a maintenance cycle for a production system consisting of $m = 3$ set-up activities and $n = 3$ components, where component j requires all set-up activities $i \leq j$ to be carried out in advance.

								components						3
		2		3	2			2	3		2			2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
				3				*set-ups*	3					3
		2		2	2			2	2		2			2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
								periods						5
		3		5	3			3	5		3			3
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
								opportunities						
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

3.2.2 Problem definition

The coordination of preventive maintenance activities is now modelled as follows: preventive maintenance on component $j \in \mathcal{J}$ is carried out at integer multiples of $k_j \cdot t$ time units, i.e. at times $\{0, k_j \cdot t, 2 \cdot k_j \cdot t, ...\}$, where $k_j \in \mathbb{N}^*$ denotes the **maintenance period** of component $j \in \mathcal{J}$ relative to a **basis interval** of $t > 0$ time units. Obviously, this corresponds to a repetitive, periodic preventive **maintenance cycle** of $\text{lcm}(k_1, ..., k_n) \cdot t$ time units, where $\text{lcm}(k_1, ..., k_n)$ denotes the least common multiplier of the integers k_j ($j \in \mathcal{J}$), e.g. $\text{lcm}(2, 3, 4) = 12$. In general, this leads to an optimization in $n + 1$ variables.

As a starting point of our analysis, we will restrict ourselves to the case where the maintenance interval $t > 0$ is fixed, and the integer periods k_j must be chosen from a finite set of possibilities, say $k_j \in \mathcal{K}$ for all $j \in \mathcal{J}$. This situation arises if t is a natural time limit (e.g. 1 day, week, or month) and one is interested in a maintenance cycle of prescribed length (e.g. 1 year). In general, this yields an optimization problem in n variables $(k_1, ..., k_n)$, with $|\mathcal{K}|^n$ different solutions, where our objective is to minimize average maintenance costs per unit of time:

$$(P) \quad \min_{k_j \in \mathcal{K}} \left\{ \sum_{i \in \mathcal{I}} \frac{s_i \cdot \Delta_i(\mathbf{k})}{t} + \sum_{j \in \mathcal{J}} \Phi_j(k_j \cdot t) \right\}.$$

Here, $\Delta_i(\mathbf{k})$ represents the fraction of times that set-up activity $i \in \mathcal{I}$ is carried out at an **opportunity** for preventive maintenance (e.g. $\Delta_1(\mathbf{k}) = 1$, $\Delta_2(\mathbf{k}) = 7/15$,

and $\Delta_3(\mathbf{k}) = 1/5$ in Table 3.1). Since each occurence of component $j \in \mathcal{J}$ implies the occurence of all set-up activities $i \in I_j$, it is easily observed that $\max\{k_j^{-1} \mid j \in J_i\} \leq \Delta_i(\mathbf{k}) \leq 1$ for all $i \in \mathcal{I}$, implying that $\Delta_i(\mathbf{k}) = 1$ if $k_j = 1$ for some $j \in J_i$. Based on the principle of inclusion and exclusion, Dagpunar (1982) derives the following general expression for $\Delta_i(\mathbf{k})$:

$$\Delta_i(\mathbf{k}) = \sum_{l=1}^{|J_i|} (-1)^{l+1} \sum_{\{j_1,...,j_l\} \subseteq J_i} \text{lcm}(k_{j_1}, ..., k_{j_l})^{-1}.$$

Typically, the basis maintenance interval t is restricted to several days or weeks, whereas the corresponding maintenance cycle $\text{lcm}(k_1, ..., k_n) \cdot t$ varies from several weeks to several months, or even years. In case of a prescribed cycle length of $T = N \cdot t > 0$ time units, appropiate values for k_j must satisfy $N \bmod k_j = 0$. Amongst others, this approach is particularly useful in calendar-based maintenance planning systems, by which workload and capacity profiles have to be matched on a regular basis. On the other hand, if there are no explicit constraints on the length of the basis maintenance interval and the corresponding maintenance cycle, our optimization problem becomes even more complex:

$$(Q) \quad \min_{t>0} \min_{k_j \in \mathbb{N}^*} \left\{ \sum_{i \in \mathcal{I}} \frac{s_i \cdot \Delta_i(\mathbf{k})}{t} + \sum_{j \in \mathcal{J}} \Phi_j(k_j \cdot t) \right\}.$$

Examples of this type can be found in various industries, e.g. if the initiation of maintenance activities is based on cumulative operating time rather than calendar moments. In such cases, the need for well-defined maintenance intervals (e.g. multiples of 100 or 1000 running hours) is less restricted, and often based on intuitive or administrative reasoning only.

3.2.3 Literature review

Up to our knowledge, problems (P) and (Q) have not been addressed in existing literature in this general form, which allows for multiple set-up activities, multiple components, and a large class of preventive maintenance models for each component. Only Gertsbakh (1977) considers a somewhat similar but less powerful modelling framework for problem (P), in which maintenance jobs can only be attached to lowest-level set-up activities, and the set of maintenance periods is of a very special structure $\mathcal{K} = \{k_1, ..., k_p\}$, where $k_1 < ... < k_p$ and k_j is an integer multiple of k_i for all $i < j$. Given this restrictive modelling framework, the author provides an efficient algorithm with which problem (P) can be solved to optimality.

On the other hand, the single set-up version ($m = 1$) of these problems has gained considerably more attention in existing literature, e.g. see Goyal and Kusy (1985), Goyal and Gunasekaran (1992), and Wildeman (1996). For computational reasons, the **correction factors** $\Delta_i(\mathbf{k})$ have usually been neglected, or - equivalently - assumed to be equal to one in the optimal strategy. Although this approach is certainly defendable in case of a single set-up activity, it is obvious that this assumption cannot be sustained within our setting of multiple set-up activities. After all, the possibility that $\Delta_i(\mathbf{k}) < 1$ for some $i \in \mathcal{I}$ is essential in our approach.

In general, finding an optimal maintenance strategy with correction factors is a very complex problem, even in case of a single set-up activity (Goyal 1982). In existing literature, we did not find any optimal or near-optimal methods to solve the single set-up versions of problems (P) and (Q) with correction factors, possibly because of the inherent mathematical complications. Even for the joint replenishment problem, where the correction factor was introduced by Dagpunar (1982), no references could be found in which a (heuristic) solution approach to these problems is presented.

3.2.4 Outline

This chapter is organized as follows. In section 3.3, we will show that problem (P) can be solved to optimality by means of a dynamic programming algorithm, as well as a mixed integer linear programming formulation. Subsequently, a lower bound for problem (Q) is derived in section 3.4, based upon which two heuristic approaches for problem (Q) are developed in sections 3.5, 3.6 and 3.7 respectively. In section 3.8, these heuristics are illustrated by means of a numerical example, which offers useful insights. Subsequently, computational results in section 3.9 show that near-optimal solutions are obtained within reasonable computation times. Finally, some conclusions and recommendations are summarized in section 3.10.

3.3 Optimization of problem (P)

As a starting point, we discuss how problem (P) can be solved by means of a dynamic programming algorithm. Next, a mixed integer linear programming formulation will be presented, with which problem (P) can also be solved to optimality. For notational convenience, a less efficient but more insightful and didactical version is presented first. Subsequently, two reduction techniques are presented which strongly reduce the size of this problem, especially if the set of possible maintenance periods $\mathcal{K}$ is large.

3.3.1 A dynamic programming algorithm

As in the previous chapter, the main underlying observation behind our dynamic programming approach is that problem (P) can be interpreted as the assignment of set-up activities $i \in \mathcal{I}$ and maintenance jobs $j \in \mathcal{J}$ to maintenance periods $k \in \mathcal{K}$, in such a way that total (expected) costs per unit of time are minimized. This is a potentially valuable insight, since the assignment of component $j \in \mathcal{J}$ to maintenance periods $k \in K$ requires all parental set-up activities $i \in I_j$ to be assigned to the same maintenance period too.

To continue our analysis, let us denote with $g_i(K)$ the minimal costs for the subtree associated with set-up activity $i \in \mathcal{I}$, provided that set-up activities and maintenance jobs within this subtree can only be assigned to maintenance periods $k \in K \subseteq \mathcal{K}$. For each set-up activity $i \in \mathcal{I}$, and each possible state $K \subseteq \mathcal{K}$, we must now decide which maintenance periods $L \subseteq K$ to use. Based upon this decision, the (optimal) assignment of maintenance jobs $j \in J_i^*$ to maintenance periods $l \in L$, as well as the consequences for all lower-level set-up activities $i' \in S_i^*$, are immediately clear. More specifically, $g_i(K)$ can be determined recursively by means of the following dynamic programming equation:

$$g_i(K) = \min_{L \subseteq K : L \neq \varnothing} \left\{ h_i(L) + \sum_{i' \in S_i^*} g_{i'}(L) \right\}.$$

Here, $h_i(L)$ denotes the (minimal) costs associated with the assignment of set-up activity $i \in \mathcal{I}$ and maintenance jobs $j \in J_i^*$ to maintenance periods $l \in L$. It is easily verified that $h_i(L)$ can be determined as follows:

$$h_i(L) = \frac{s_i \cdot \Delta_i(L)}{t} + \sum_{j \in J_i^*} \min_{l_j \in L} \{\Phi_j(l_j \cdot t)\}.$$

Under some weak conditions, the calculation of $h_i(L)$ can be performed in a rather straightforward manner. We will come back to that later on in this chapter. Here, we only mention that the optimal solution to problem (P) can now be determined recursively by calculation of $g_1(\mathcal{K})$, since $i = 1$ denotes the common i.e. highest-level set-up activity. In the worst case, this yields a dynamic programming algorithm with $2^{|\mathcal{K}|}$ states and decisions, where $|\mathcal{K}|$ denotes the number of different maintenance frequencies.

On the other hand, the set of relevant states $K \subseteq \mathcal{K}$ and decisions $L \subseteq K$ for set-up activity $i \in \mathcal{I}$ can often be reduced significantly, by observing that $h_i(L) \geq h_i(L')$ for some $L \subset L' \subseteq K$ implies that decision $L \subseteq K$ can as well be neglected. After all,

it leads to higher (direct) costs and less flexibility for all lower-level set-up activities. At the very least, this means that $\Delta_i(L) = \Delta_i(L')$ for some $L \subset L' \subseteq K$ implies that decision $L \subseteq K$ can as well be neglected. If $K = \{1, 2, 3, 4\}$, this means that the only relevant decisions are $\{1, 2, 3, 4\}$, $\{2, 3, 4\}$, $\{2, 4\}$, $\{3, 4\}$, $\{3\}$ and $\{4\}$. All other decisions can as well be neglected, since e.g. $\Delta(\{2, 3\}) = \frac{4}{6} = \frac{8}{12} = \Delta(\{2, 3, 4\})$ and $\{2, 3\} \subseteq \{2, 3, 4\}$.

Obviously, these nice structural properties should be further exploited in formulating an efficient dynamic programming algorithm for the optimization of problem (P). Nevertheless, this approach would still become inattractive, or even intractable, if the number of maintenance periods grows too large (e.g. $|\mathcal{K}| > 10$). In such cases, it is also possible, and probably more efficient to solve the problem by means of a mixed integer linear programming formulation. In the following section, this alternative approach will be discussed in more detail.

3.3.2 A mixed integer linear programming formulation

Let us now present a mixed integer linear programming formulation, with which problem (P) can also be solved to optimality. The underlying observation behind this formulation is that, given the basis maintenance interval of $t > 0$ time units, the assignment of components $j \in \mathcal{J}$ to maintenance periods $k \in \mathcal{K}$ will always result in a preventive maintenance cycle of at most $\text{lcm}(\mathcal{K}) \cdot t$ time units. In line with this, we denote with $\mathcal{L} = \{1, ..., \text{lcm}(\mathcal{K})\}$ the set of so-called maintenance opportunities $\{0, t, 2 \cdot t, ...\}$ within this preventive maintenance cycle (e.g. $|\mathcal{L}| = 15$ in Table 3.1).

To continue our analysis, we denote with $K_l = \{k \in \mathcal{K} \mid l \bmod k = 0\}$ the set of maintenance periods that correspond with maintenance opportunity $l \in \mathcal{L}$ (e.g. $K_5 = K_{10} = \{1, 5\}$ in Table 3.1). The problem now consists of assigning components $j \in \mathcal{J}$ to maintenance periods $k \in \mathcal{K}$, in such a way that the costs of the corresponding maintenance cycle are minimized. In our model, the assignment of set-up activities $i \in \mathcal{I}$ to maintenance opportunities $l \in \mathcal{L}$ is comprised into variables $x_{il} \in \{0, 1\}$, whereas the assignment of components $j \in \mathcal{J}$ to maintenance periods $k \in \mathcal{K}$ is represented by variables $y_{jk} \in \{0, 1\}$:

$$x_{il} = \begin{cases} 1 & \text{if set-up } i \in \mathcal{I} \text{ is assigned to opportunity } l \in \mathcal{L}, \\ 0 & \text{otherwise,} \end{cases}$$

$$y_{jk} = \begin{cases} 1 & \text{if component } j \in \mathcal{J} \text{ is assigned to period } k \in \mathcal{K}, \\ 0 & \text{otherwise.} \end{cases}$$

With $a_{il} = s_i/(\text{lcm}(\mathcal{K}) \cdot t)$, we denote the average cost per unit of time associated with the assignment of set-up activity $i \in \mathcal{I}$ to maintenance opportunity $l \in \mathcal{L}$. Similarly, $b_{jk} = \Phi_j(k \cdot t)$ denotes the average cost per unit of time associated with the assignment of component $j \in \mathcal{J}$ to maintenance period $k \in \mathcal{K}$. With this in mind, problem (P) can now be formulated in terms of a mixed integer linear program, where our objective is to minimize average maintenance costs per unit of time:

$$\text{Minimize} \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}} a_{il} \cdot x_{il} + \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} b_{jk} \cdot y_{jk}$$

Subject to:

$$\begin{array}{lrcll}
\text{(a)} & x_{il} & \geq & x_{i'l} & \forall i \in \mathcal{I}, l \in \mathcal{L}, i' \in S_i^* \\
\text{(b)} & x_{il} & \geq & y_{jk} & \forall i \in \mathcal{I}, l \in \mathcal{L}, j \in J_i^*, k \in K_l \\
\text{(c)} & \sum_{k \in \mathcal{K}} y_{jk} & = & 1 & \forall j \in \mathcal{J} \\
\text{(d)} & x_{il} & \geq & 0 & \forall i \in \mathcal{I}, l \in \mathcal{L} \\
\text{(e)} & y_{jk} & \in & \{0,1\} & \forall j \in \mathcal{J}, k \in \mathcal{K}
\end{array}$$

Here, restriction (a) states that the assignment of set-up activity $i \in \mathcal{I}$ to maintenance opportunity $l \in \mathcal{L}$ requires the parental set-up activities to be carried out at the same maintenance opportunity too. Similarly, restriction (b) ensures that the assignment of component $j \in \mathcal{J}$ to maintenance period $k \in \mathcal{K}$ causes the corresponding set-up activity to be carried out at the corresponding maintenance opportunities too. Moreover, restrictions (c) and (e) guarantee that exactly one period is assigned to each component. As a consequence, restriction (d) is sufficient to ensure that $x_{il} \in \{0,1\}$ for all $i \in \mathcal{I}$ and $l \in \mathcal{L}$.

3.3.3 Problem reduction

In general, the number of maintenance opportunities $\text{lcm}(\mathcal{K})$ grows exponentially with the set of possible maintenance periods $\mathcal{K}$. Therefore, we have developed a more efficient mixed integer linear programming formulation, which is based on the following observations:

- maintenance opportunities $l \in \mathcal{L}$ with $K_l = \varnothing$ can as well be left out of consideration, since evidently $x_{il} = 0$ for all $i \in \mathcal{I}$ in an optimal solution;
- maintenance opportunities $l_1, l_2 \in \mathcal{L}$, with $K_{l_1} = K_{l_2}$ and $l_1 \neq l_2$, can as well be replaced by a single maintenance opportunity l with $K_l = K_{l_1} = K_{l_2}$ and corresponding cost $a_{il} = a_{l_1} + a_{l_2}$ for all $i \in \mathcal{I}$, since evidently $x_{il_1} = x_{il_2}$ for all $i \in \mathcal{I}$ in an optimal solution.

Obviously, these observations may lead to significant reductions in the problem size. To be specific, let us denote with $\mathcal{L}^*$ the reduced set of maintenance opportunities, and with $\eta(U)$ the number of times that precisely cluster $U \subseteq \mathcal{K}$ shows up in the maintenance cycle (e.g. $\eta(\{1,3\}) = 4$ and $\eta(\{1,5\}) = 2$ in Table 3.1). Then the number of opportunities $|\mathcal{L}^*|$ to be considered in the reduced version of the problem equals the number of clusters $U \subseteq \mathcal{K}$ with $\eta(U) > 0$. Based on the principle of inclusion and exclusion, $\eta(U)$ yields an expression which is similar to $\Delta_i(\mathbf{k})$:

$$\eta(U) = \text{lcm}(\mathcal{K}) \cdot \sum_{l=|U|}^{|\mathcal{K}|} (-1)^{l-|U|} \sum_{U \subseteq \{k_1,...,k_l\} \subseteq \mathcal{K}} \text{lcm}(k_1, ..., k_l\}^{-1}.$$

In general, the determination of $\eta(U)$ with the use of this equation is a complex problem, since the number of clusters $U \subseteq \mathcal{K}$ grows exponentially with the number of possible maintenance periods $\mathcal{K}$. For similar reasons, enumeration of all maintenance opportunities up to $\text{lcm}(\mathcal{K})$ is no valid option. Apparently, a more efficient method has to be constructed. In this respect, it is useful to observe that the first occurence (if any) of cluster $U \subseteq \mathcal{K}$ in the maintenance cycle must be observed at maintenance opportunity $l = \text{lcm}(U)$. This means that the following relation holds for all $U \subseteq \mathcal{K}$:

$$\eta(U) > 0 \Longleftrightarrow \text{lcm}(U) \bmod k \neq 0 \text{ for all } k \in \mathcal{K} \setminus U.$$

By doing this, a large number of clusters can usually be discarded beforehand. Moreover, the frequency of occurence of the remaining clusters within the maintenance cycle can be evaluated by means of the following implicit relation, which holds for all $U \subseteq \mathcal{K}$, and follows simply by taking $\text{lcm}(U)$ as the new time step. Here, we define $\text{lcm}(\varnothing) = 1$ for notational convenience:

$$\text{lcm}(U) \cdot \sum_{V \supseteq U} \eta(V) = \text{lcm}(\mathcal{K}).$$

3.3.4 An iterative procedure to construct $\mathcal{L}^*$

Our analysis now proceeds as follows. As a starting point, the set of maintenance periods $\mathcal{K} = \{k_1, ..., k_p\}$ is ordered such that $k_1 < ... < k_p$. Subsequently, we define Ω_i as the set of clusters $U \subseteq \{k_1, ..., k_i\}$ with $\eta(U) > 0$, in case the set of possible maintenance periods would be reduced to $\{k_1, ..., k_i\}$. Obviously, we have $\Omega_1 = \{\{1\}\}$ if $k_1 = 1$, and $\Omega_1 = \{\varnothing, \{k_1\}\}$ otherwise. Since we are mainly interested in Ω_p, it is now sufficient to formulate an (efficient) iterative procedure which constructs Ω_{i+1} from Ω_i.

First of all, we observe that $U \in \Omega_{i+1}$ implies that either $U \in \Omega_i$ or $U \setminus \{k_{i+1}\} \in \Omega_i$, or both. Hence, it is sufficient to determine for each $U \in \Omega_i$ whether, and under which conditions $U \in \Omega_{i+1}$ and/or $U \cup \{k_{i+1}\} \in \Omega_{i+1}$, in order to construct Ω_{i+1} from Ω_i. It is easily verified that this can be done in the following, rather straightforward way. Starting with $\Omega_{i+1} = \varnothing$, it is determined for each $U \in \Omega_i$ whether either one, or both of the following conditions are satisfied:

(i) $\text{lcm}(U) \bmod k_{i+1} \neq 0$,

(ii) $\text{lcm}(U \cup \{k_{i+1}\}) \bmod k \neq 0$ for all $k \in \{k_1, ..., k_i\} \setminus U$.

If condition (i) is satisfied, Ω_{i+1} is extended with cluster U. Moreover, if condition (ii) is satisfied, Ω_{i+1} is extended with cluster $U \cup \{k_{i+1}\}$. The underlying observation behind these conditions is that the first occurence (if any) of cluster $U \subseteq \{k_1, ..., k_{i+1}\}$ must take place at maintenance opportunity $l = \text{lcm}(U)$. But this occurs if and only if $\text{lcm}(U) \bmod k \neq 0$ for all $k \in \{k_1, ..., k_{i+1}\} \setminus U$. Note that condition (ii) is always true if $U = \{k_1, ..., k_i\}$, and thus $\{k_1, ..., k_{i+1}\} \in \Omega_{i+1}$ for all $i < p$. Once Ω_p has been determined, the values of $\eta(U)$ for all $U \in \Omega_p$ are determined recursively as follows:

$$\eta(U) = \frac{\text{lcm}(\mathcal{K})}{\text{lcm}(U)} - \sum_{V \in \Omega_p : U \subset V} \eta(V).$$

As an example, consider the maintenance cycle of Table 3.1. Then it is easily verified that $\Omega_1 = \{\{1\}\}$, $\Omega_2 = \{\{1\}, \{1,3\}\}$, and $\Omega_3 = \{\{1\}, \{1,3\}, \{1,5\}, \{1,3,5\}\}$. Starting with $\eta(\{1,3,5\}) = 1$, the remaining clusters are evaluated as follows:

$$\begin{aligned} \eta(\{1,3\}) &= \text{lcm}(1,3,5)/\,\text{lcm}(1,3) - \eta(\{1,3,5\}) = 4, \\ \eta(\{1,5\}) &= \text{lcm}(1,3,5)/\,\text{lcm}(1,5) - \eta(\{1,3,5\}) = 2, \\ \eta(\{1\}) &= \text{lcm}(1,3,5)/\,\text{lcm}(1) - \eta(\{1,3\}) - \eta(\{1,5\}) - \eta(\{1,3,5\}) = 8. \end{aligned}$$

Clearly, this approach yields an alternative problem formulation with significantly less maintenance opportunities, and thus variables and constraints (see Table 3.2). But even in the reduced version of our problem, the number of maintenance opportunities may still grow exponentially with the number of maintenance periods. On the other hand, it is also possible that the number of maintenance opportunities equals the number of maintenance periods, i.e. $|\mathcal{L}^*| = |\mathcal{K}|$. An example of this type was presented by Gertsbakh (1977), who considers the case where each possible maintenance period is an integer multiple of all other, but smaller maintenance periods. Although this problem can be solved efficiently with the use of dynamic programming, our more general mixed integer linear programming formulation is also well-equiped to deal with such problems, due to the small amount of maintenance opportunities.

Table 3.2: Reductions in problem size for several period sets $\mathcal{K} = \{1, ..., n\}$.

n	1	2	3	5	10	15	25
$\lvert\mathcal{L}\rvert$	1	2	6	60	2520	360360	26771144400
$\lvert\mathcal{L}^*\rvert$	1	2	4	12	48	192	2880

3.3.5 Numerical validation

Of course, our mixed integer linear programming formulation was tested on a series of randomly created (small) test problems. For a detailed description of these test problems, we refer to the computational results in section 3.9. Here, we only mention that the LP-relaxation generated an integer, and thus feasible solution to problem (P), in approximately 68% of all test problems, whereas the average deviation with respect to the optimal solution was only 0.89% in all other cases. Moreover, an optimal integer solution was usually found within a few iterations of the consecutive branch and bound algorithm. Recall that similar, and even stronger results were observed for the clustering problem with frequency-constrained maintenance jobs, as considered in the previous chapter. Probably, this phenomenon originates from the fact that problem (P) is closely related to a standard assignment problem, which is known to possess the above-mentioned integrality property. Apparently, the underlying hierarchical structure of set-up activities and components does not conflict too much with this nice and useful property of the standard assignment problem.

3.4 A lower bound for problem (Q)

In general, problem (Q) is a complex mixed continuous-integer programming problem, which makes it difficult to solve to optimality. Under some weak conditions, however, its complexity can be reduced significantly. In this section, we will present a relaxation of problem (Q), with which a lower bound for the optimal solution can be derived. To a certain extent, this approach is based upon, and therefore similar to Wildeman (1996), who developed an optimal solution approach to the single set-up version of this problem without correction factors. Nevertheless, it contains a variety of interesting new elements which are typical for our setting of multiple set-up activities.

3.4.1 Model assumptions

As a starting point of our analysis, we assume that $M_j(x)$ is (i) strictly positive, (ii) strictly increasing, (iii) strictly convex and (iv) twice continuously differentiable on

$\langle 0, \infty \rangle$ for all $j \in \mathcal{J}$. Simply stated, these assumptions are related to an increasing marginal cost rate for postponing preventive maintenance activities (Berg 1980). In general, they account for a large class of preventive maintenance models of the block type, including the minimal repair model (Barlow and Hunter 1960), the standard inspection model (Barlow, Hunter, and Proschan 1963), and the delay time model (Christer and Waller 1984). On the other hand, if one or more components are modelled according to a standard block replacement model (Barlow and Proschan 1965) or modified block replacement model (Berg and Epstein 1976), assumption (iii) may not be satisfied. In that case, however, our methods can still be used to obtain approximate results.

Lemma 2 *Suppose that $M_j(x)$ is strictly positive, strictly increasing, strictly convex and twice continuously differentiable on $\langle 0, \infty \rangle$. Then $\Phi_j(x)$ has exactly one local minimum $x_j^* < \infty$. Moreover, $\Phi_j(1/x)$ is strictly convex on $\langle 0, \infty \rangle$.*

Proof. Let us first derive some algebraic expressions for the first and second derivatives of the deterioration cost function $\Phi_j(x) = (c_j + M_j(x))/x$, for $x > 0$:

$$\Phi_j'(x) = \frac{M_j'(x) \cdot x - (c_j + M_j(x))}{x^2} = \frac{M_j'(x) - \Phi_j(x)}{x},$$

$$\Phi_j''(x) = \frac{(M_j''(x) - \Phi_j'(x)) \cdot x - (M_j'(x) - \Phi_j(x))}{x^2} = \frac{M_j''(x) - 2 \cdot \Phi_j'(x)}{x}.$$

Since each local minimum $x_j^* > 0$ must at least satisfy $\Phi_j'(x_j^*) = 0$ and $\Phi_j''(x_j^*) \geq 0$, it follows from the above equations that these conditions can be formulated equivalently as $M_j'(x_j^*) = \Phi_j(x_j^*)$ and $M_j''(x_j^*) \geq 0$. Since $\Phi_j(x_j^*) < \infty$ by definition, and $M_j(x)$ is assumed to be strictly increasing and strictly convex, this implies the uniqueness of x_j^*. For similar reasons, $\lim_{x \to \infty} M_j'(x) = \infty$ yields $x_j^* < \infty$. Finally, with $\Phi_j(1/x) = x \cdot (c_j + M_j(1/x))$, it is easily verified after some elementary algebra that $\frac{\partial^2}{\partial x^2} \Phi_j(1/x) = M_j''(1/x)/x^3$. Since $M_j''(1/x) > 0$ for all $x > 0$ by assumption, this implies that $\Phi_j(1/x)$ is strictly convex on $\langle 0, \infty \rangle$. $\square$

3.4.2 Problem relaxation

In view of Lemma 1, it seems reasonable to rewrite problem (Q) by transformation of t into t^{-1}. By doing this, the objective function becomes a convex function in t, which obviously is a useful property in finding the optimal preventive maintenance cycle. In this alternative formulation, preventive maintenance on component $j \in \mathcal{J}$

is carried out at integer multiples of k_j/t time units. Here, $k_j \in \mathbb{N}^*$ denotes the maintenance period of component $j \in \mathcal{J}$ relative to a basis interval of $t^{-1} > 0$ time units:

$$(Q) \quad \min_{t>0} \min_{k_j \in \mathbb{N}^*} \left\{ \sum_{i \in \mathcal{I}} s_i \cdot \Delta_i(\mathbf{k}) \cdot t + \sum_{j \in \mathcal{J}} \Phi_j(k_j/t) \right\}.$$

Let us now present a relaxation of problem (Q), which enables us to construct a lower bound for the optimal solution with the use of standard search techniques. As a starting point, define scalars $0 \leq \alpha_{ij} \leq 1$, such that $\sum_{i \in I_j} \alpha_{ij} = 1$ for all $j \in \mathcal{J}$. Here, α_{ij} could be interpreted as the contribution of set-up activity $i \in I_j$ to the costs of component $j \in \mathcal{J}$. Now, observe that the following relation holds:

$$\sum_{j \in \mathcal{J}} \Phi_j(k_j/t) = \sum_{j \in \mathcal{J}} \sum_{i \in I_j} \alpha_{ij} \cdot \Phi_j(k_j/t) = \sum_{i \in \mathcal{I}} \sum_{j \in J_i} \alpha_{ij} \cdot \Phi_j(k_j/t).$$

With $\mathcal{A} \subseteq \mathbb{R}_+^{mn}$, we denote the set of all such weights $(\alpha_{11}, ..., \alpha_{mn})$. For each $\boldsymbol{\alpha} \in \mathcal{A}$, this yields an alternative formulation for problem (Q), in which the individual costs of each component are divided among the corresponding set-up activities:

$$\min_{t>0} \min_{k_j \in \mathbb{N}^*} \sum_{i \in \mathcal{I}} \left\{ s_i \cdot \Delta_i(\mathbf{k}) \cdot t + \sum_{j \in J_i} \alpha_{ij} \cdot \Phi_j(k_j/t) \right\}.$$

To continue our analysis, we substitute $t_i = \Delta_i(\mathbf{k}) \cdot t$ and observe that $t_i \geq t_{i'}$ if $i' \in S_i$, since obviously $\Delta_i(\mathbf{k}) \geq \Delta_{i'}(\mathbf{k})$ if set-up activity $i \in \mathcal{I}$ is a parent of set-up activity $i' \in \mathcal{I}$. If we denote with $\mathcal{T} \subseteq \mathbb{R}_+^m$ the set of feasible solutions to $(t_1, ..., t_m)$, this yields the following lower bound for problem (Q):

$$\min_{\mathbf{t} \in \mathcal{T}} \min_{k_j \in \mathbb{N}^*} \sum_{i \in \mathcal{I}} \left\{ s_i \cdot t_i + \sum_{j \in J_i} \alpha_{ij} \cdot \Phi_j(k_j \cdot \Delta_i(\mathbf{k})/t_i) \right\}.$$

Now we substitute $k_{ij} = k_j \cdot \Delta_i(\mathbf{k})$ and observe that $k_{ij} \geq 1$ for all $i \in \mathcal{I}$ and $j \in J_i$, since $\Delta_i(\mathbf{k}) \geq k_j^{-1}$ for all $j \in J_i$ by definition. As before, this yields an alternative lower bound for problem (Q):

$$\min_{\mathbf{t} \in \mathcal{T}} \min_{k_{ij} \geq 1} \sum_{i \in \mathcal{I}} \left\{ s_i \cdot t_i + \sum_{j \in J_i} \alpha_{ij} \cdot \Phi_j(k_{ij}/t_i) \right\}.$$

For notational convenience, let us now define functions $\xi_j(t) = \min\{\Phi_j(x/t) \mid x \geq 1\}$ for all $j \in \mathcal{J}$. Since $\Phi_j(x)$ attains its minimum at $x_j^* < \infty$, and $\Phi_j(1/x)$ is strictly convex on $\langle 0, \infty \rangle$, this yields a decreasing function in t, which is convex on $\langle 0, \infty \rangle$, and strictly convex on $\langle 0, 1/x_j^* \rangle$:

$$\xi_j(t) = \min_{x \geq 1} \{\Phi_j(x/t)\} = \begin{cases} \Phi_j(1/t) & \text{if } t \leq 1/x_j^* \\ \Phi_j(x_j^*) & \text{if } t \geq 1/x_j^* \end{cases}$$

Since we are free to choose $\boldsymbol{\alpha} \in \mathcal{A}$ arbitrarily, this finally leaves us with the following lower bound for problem (Q). Obviously, this lower bound reflects a decomposition into m interrelated subproblems, each corresponding with a single set-up activity, and its adjacent components:

$$\min_{\mathbf{t} \in \mathcal{T}} \max_{\boldsymbol{\alpha} \in \mathcal{A}} \sum_{i \in \mathcal{I}} \left\{ s_i \cdot t_i + \sum_{j \in J_i} \alpha_{ij} \cdot \xi_j(t_i) \right\}.$$

Our analysis now proceeds as follows. Since $\xi_j(t)$ is a decreasing function in t for all $j \in \mathcal{J}$, and $t_i \geq t_{i'}$ if $i' \in S_i$, it can easily be verified that the optimal values of α_{ij} are determined as follows: $\alpha_{ij} = 1$ if $j \in J_i^*$, and $\alpha_{ij} = 0$ otherwise. Recall that $J_i^* \subseteq J_i$ denotes the set of components $j \in \mathcal{J}$ that are attached to set-up activity $i \in \mathcal{I}$. Summarizing, this yields the following lower bound (R) for problem (Q):

$$(R) \quad \min_{\mathbf{t} \in \mathcal{T}} \sum_{i \in \mathcal{I}} \left\{ s_i \cdot t_i + \sum_{j \in J_i^*} \xi_j(t_i) \right\}.$$

Since $s_i \cdot t_i$ and $\xi_j(t_i)$ are both convex functions in t_i, and $J_i \neq \varnothing$ for all $i \in \mathcal{I}$, this leaves us with a convex programming problem in a convex search space $\mathcal{T} \subseteq \mathbb{R}_+^m$, which can easily be solved to optimality with the use of standard search techniques. Here, we applied the gradient projection method (Luenberger 1984). To this end, we used the following initial values $(\theta_1^*, ..., \theta_m^*)$ for $(t_1^*, ..., t_m^*)$ in deriving the optimal solution to problem (R):

$$\theta_i^* = \max_{j \in J_i} \left\{ \frac{1}{x_j^*} \right\}.$$

It is easily verified that $(\theta_1^*, ..., \theta_m^*) \in \mathcal{T}$, since obviously $i' \in S_i$ implies that $J_{i'} \subseteq J_i$, and thus $\theta_{i'}^* \leq \theta_i^*$. Of course, other initial values might be preferred, e.g. if they require a smaller number of iterations of the underlying gradient projection method. Nevertheless, they will always result in the same lower bound, as is more stated explicitly in the following theorem:

Lemma 3 *There exists a unique solution* $(t_1^*, ..., t_m^*) \leq (\theta_1^*, ..., \theta_m^*)$ *to problem* (R).

Proof. First of all, suppose that $t_i^* > \theta_i^*$ for some $i \in \mathcal{I}$, or equivalently $t_i^* > 1/x_j^*$ for all $j \in J_i$. Without loss of generality, we can assume that $t_{i'}^* \leq \theta_{i'}^*$ for all $i' \in S_i$. Since $s_i \cdot \theta_i^* < s_i \cdot t_i^*$, and $\xi_j(\theta_i^*) = \xi_j(t_i^*) = \Phi_j(x_j^*)$ for all $j \in J_i$, it is now immediately clear that $(t_1^*, ..., t_i^*, ..., t_m^*)$ is outperformed by $(t_1^*, ..., \theta_i^*, ..., t_m^*)$, a contradiction. Apparently, $t_i^* > \theta_i^*$ cannot be optimal, and thus $t_i^* \leq \theta_i^*$ for all $i \in \mathcal{I}$. Because $s_i \cdot t_i + \sum_{j \in J_i^*} \xi_j(t_i)$ is strictly convex in t_i for all $i \in \mathcal{I}$, this also implies the uniqueness of $(t_1^*, ..., t_m^*)$, which completes the proof. □

3.5 An iterative heuristic for problem (Q)

In this section, we will present an iterative heuristic approach to solve problem (Q). This approach is based on decomposition of problem (Q) into two subproblems: one that determines $(k_1, ..., k_n)$ given t, denoted (Q_k), and one that determines t given $(k_1, ..., k_n)$, denoted (Q_t). As a starting point, however, we determine the optimal solution $(t_1^*, ..., t_m^*)$ to problem (R), and initialize $t = t_1^*$. Subsequently, subproblems (Q_k) and (Q_t) are solved iteratively, until no improvements are observed in two consecutive iterations (see Figure 3.2). To a certain extent, this approach is similar to Goyal and Kusy (1985) and Goyal and Gunasekaran (1992), who developed near-optimal solutions to the single set-up version of problem (Q) without correction factor. Nevertheless, it contains a variety of interesting elements which are typical for our setting of multiple set-up activities with correction factors.

Simply stated, the initial choice $t = t_1^*$ is based on the intuitive reasoning that $\Delta_1(\mathbf{k}) = 1$ in the optimal maintenance cycle, which on its turn implies that $k_j = 1$ for at least one component $j \in \mathcal{J}$. From a practical point of view, this relates to a preventive maintenance cycle without empty maintenance opportunities. Of course, other initial values for t could lead to other, and possibly better solutions. Nevertheless, a series of numerical experiments carried out by Wildeman (1996) strongly indicate that this approach is a promising one, at least in the single set-up version of our problem $(m = 1)$. In the following sections, we will show how problem (Q_t) can be solved to optimality with the use of standard search techniques. Moreover, we will present two heuristics for problem (Q_k).

3.6 Optimization of problem (Q_t)

If the (optimal) maintenance periods $(k_1, ..., k_n)$ are known, it is possible to compute the values of $\Delta_i(\mathbf{k})$ beforehand for all $i \in \mathcal{I}$. To be specific, if we denote with $\eta_i(\emptyset)$

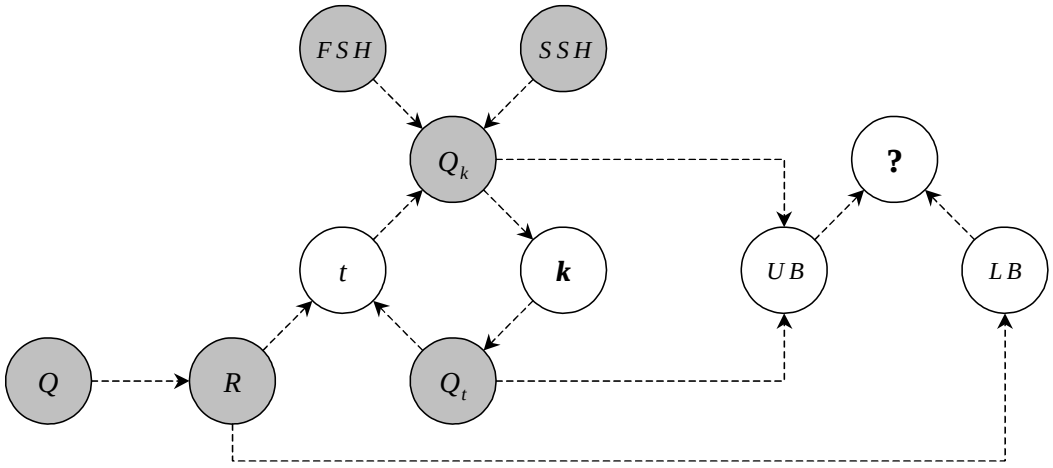


Figure 3.2: Contribution of subproblems (R), (Q_t) and (Q_k), as well as heuristics (FSH) and (SSH), to the derivation of a lower bound (LB) and upper bound (UB) for the optimal solution to problem (Q).

the number of empty maintenance opportunities $l \in \mathcal{L}$ with $K_l = \emptyset$ in a maintenance cycle with period set $\mathcal{K}_i = \{k_j \mid j \in J_i\}$, it is easily verified that the following relation holds:

$$\Delta_i(\mathbf{k}) = 1 - \frac{\eta_i(\emptyset)}{\text{lcm}(\mathcal{K}_i)}$$

After all, the maintenance cycle for set-up activity $i \in \mathcal{I}$ repeats itself after each $\text{lcm}(\mathcal{K}_i)$ maintenance opportunities, of which $\eta_i(\emptyset)$ refer to empty ones. Hence, set-up activity $i \in \mathcal{I}$ must be carried out at $\text{lcm}(\mathcal{K}_i) - \eta_i(\emptyset)$ out of $\text{lcm}(\mathcal{K}_i)$ maintenance opportunities, which yield the desired result for $\Delta_i(\mathbf{k})$. If we now denote with $\hat{s}_i = s_i \cdot \Delta_i(\mathbf{k})$ the costs associated with set-up activity $i \in \mathcal{I}$, problem (Q) reduces to the following optimization problem:

$$(Q_t) \quad \min_{t>0} \left\{ \sum_{i \in \mathcal{I}} \hat{s}_i \cdot t + \sum_{j \in \mathcal{J}} \Phi_j(k_j/t) \right\}$$

Since $\hat{s}_i \cdot t$ and $\Phi_j(k_j/t)$ are both convex in t, this leaves us with a one-dimensional convex programming problem, which can easily be solved to optimality using standard search techniques.

3.7 Heuristics for problem (Q_k)

If the (optimal) maintenance interval $1/t$ is known, problem (Q) reduces to the following optimization problem:

$$(Q_k) \quad \min_{k_j \in \mathbb{N}^*} \left\{ \sum_{i \in \mathcal{I}} s_i \cdot \Delta_i(\mathbf{k}) \cdot t + \sum_{j \in \mathcal{J}} \Phi_j(k_j/t) \right\}$$

In general, this problem is very difficult to solve, due to the unlimited choice of maintenance periods $k_j \in \mathbb{N}^*$, as well as the complex underlying structure of the correction factors $\Delta_i(\mathbf{k})$. Under some special conditions, however, the complexity of problem (Q_k) can be reduced significantly. This is particularly true for the modelling framework presented by Gertsbakh (1977), who reduces the set of possible maintenance periods to $\{a_1, a_1 \cdot a_2, ..., a_1 \cdot ... \cdot a_p\}$, where a_i $(1 \leq i \leq p)$ are all positive integers.

In this section, we will present two heuristics in which considerably more degrees of freedom are taken into account. In the first heuristic, we restrict ourselves to a finite set of consecutive maintenance periods $\mathcal{K} = \{k_{\min}, ..., k_{\max}\}$. We solve this problem to optimality by means of the mixed integer linear programming formulation, that was designed for problem (P). In the second heuristic, we consider the case where each set-up activity is carried out, and set-up costs $s_i > 0$ are incurred, at fixed intervals of $\hat{k}_i/t$ time units. Subsequently, we show that this problem can be solved to optimality by means of an efficient dynamic programming formulation. For notational convenience, we will refer to these heuristics as the finite set heuristic (FSH), and the structured set heuristic (SSH) respectively.

3.7.1 A finite set heuristic

In the finite set heuristic (FSH), each subproblem (Q_k) is solved by means of the mixed integer linear programming formulation, that was designed for problem (P). To this end, the set of possible maintenance periods is reduced to $\mathcal{K} = \{k_{min}, ..., k_{max}\}$, where k_{min} and k_{max} are determined as follows. For each component $j \in \mathcal{J}$, we determine the optimal (individual) maintenance interval $x_j^* > 0$ if no preventive set-up costs would be charged. In a similar way, we also determine the optimal (individual) maintenance interval $y_j^* > 0$ for each component $j \in \mathcal{J}$, if preventive set-up costs $\sum_{i \in I_j} s_i > 0$ would be charged:

$$x_j^* = \arg \min_{x>0} \left\{ \frac{c_j + M_j(x)}{x} \right\}$$

$$y_j^* = \arg \min_{y>0} \left\{ \frac{\sum_{i \in I_j} s_i + c_j + M_j(y)}{y} \right\}$$

The finite set heuristic is now based on the (intuitive) reasoning that $\min\{x_j^* \mid j \in \mathcal{J}\}$ and $\max\{y_j^* \mid j \in \mathcal{J}\}$ can be interpreted as a lower and upper bound on the optimal maintenance interval for each component $j \in \mathcal{J}$. More specifically, we assume that the maintenance periods $(k_1^*, ..., k_n^*)$ in each optimal solution to problem (Q_k) will satisfy:

$$k_{min} = \min_{j \in \mathcal{J}} \left\lfloor x_j^* \cdot t \right\rfloor \leq \min_{j \in \mathcal{J}} \left\{k_j^*\right\} \leq \max_{j \in \mathcal{J}} \left\{k_j^*\right\} \leq \max_{j \in \mathcal{J}} \left\lceil y_j^* \cdot t \right\rceil = k_{max}$$

As we shall see in the remainder of this chapter, we have had no indications (so far) that these bounds are too tight (see section 3.9). Clearly, we could come up with even weaker lower and upper bounds, but with respect to the efficiency of our MILP formulation, this is certainly not desirable. Since $M_j(\cdot)$ is assumed to be strictly increasing and convex on $\langle 0, \infty \rangle$, it follows from Lemma 1 that $0 < x_j^* < y_j^* < \infty$ for all $j \in \mathcal{J}$, and thus $0 \leq k_{min} \leq k_{max} < \infty$. Obviously, we do not allow $k_{min} = 0$, and in that case set $k_{min} = 1$.

It is to be expected that the finite set heuristic becomes intractable when the set of possible maintenance periods becomes too large, e.g. $\mathcal{K} = \{1, ..., 25\}$ with $|\mathcal{L}^*| = 2880$ maintenance opportunities. In that case, we suggest the use of other methods, which more explicitly account for large fluctuations in the optimal maintenance intervals x_j^* and y_j^* for different components. A typical example of this type is the so-called structured set heuristic, as will be explained in more detail in the following section.

3.7.2 A structured set heuristic

In the structured set (SSH) heuristic, we restrict ourselves to the case where each set-up activity $i \in \mathcal{I}$ is carried out, and set-up costs $s_i > 0$ are incurred, at fixed intervals of $\hat{k}_i/t > 0$ time units. Here, $\hat{k}_i \in \mathbb{N}^*$ is a new decision variable representing the maintenance period of set-up activity $i \in \mathcal{I}$. By doing so, it is immediately clear that k_j must be an integer multiple of $\hat{k}_i$ (i.e. $k_j \bmod \hat{k}_i = 0$) for all $j \in J_i$, and $\hat{k}_{i'}$ must be an integer multiple of $\hat{k}_i$ (i.e. $\hat{k}_{i'} \bmod \hat{k}_i = 0$) for all $i' \in S_i$. Therefore, an equivalent interpretation is to assume that the correction factors $\Delta_i(\mathbf{k})$ can be determined as follows:

$$\Delta_i(\mathbf{k}) = \frac{1}{\gcd\{k_j \mid j \in J_i\}}$$

Here, $\gcd(k_1, ..., k_p)$ denotes the greatest common divisor of the integers $k_1 ... k_p$, e.g. $\gcd(6, 9, 12) = 3$. It is easily verified that $\Delta_i(\mathbf{k}) \leq 1/\gcd\{k_j \mid j \in J_i\}$ under all circumstances. Hence, this expression for $\Delta_i(\mathbf{k})$ is correct in case $\gcd\{k_j \mid j \in J_i\} =$

$\min\{k_j \mid j \in J_i\}$, since obviously $\Delta_i(\mathbf{k}) \geq 1/\min\{k_j \mid j \in J_i\}$ for all $i \in \mathcal{I}$. In all other cases, it serves as an upper bound. If each maintenance period must be chosen from a finite set of possibilities $\{a_1, a_1 \cdot a_2, ..., a_1 \cdot ... \cdot a_p\}$, where all a_i $(1 \leq i \leq p)$ are positive integers, it is obvious that $\gcd\{k_j \mid j \in J_i\} = \min\{k_j \mid j \in J_i\}$ under all circumstances. In this respect, our approach provides a much richer and more powerful modelling framework in comparison with the one presented by Gertsbakh (1977). More specifically, it does so in each of the following dimensions:

- it is also possible to define components at each set-up activity in the maintenance tree, and not at the lowest-level set-up activities only;
- it does not require a set of predetermined maintenance periods to choose from, each of which is an integer multiple of all other but smaller maintenance periods;
- it always outperforms the optimal solution found by Gertsbakh (1977), irrespective of how much and which maintenance periods are used.

Now the structured-set heuristic proceeds in the following way. As a starting point, we determine the optimal solution k_j $(j \in \mathcal{J})$ to problem (Q_k) with approximate correction factors $\Delta_i(\mathbf{k}) = 1/\gcd\{k_j \mid j \in J_i\}$. To achieve this, we have developed an efficient dynamic programming formulation, which is presented next. Subsequently, we determine the actual values of $\Delta_i(\mathbf{k})$ for each set-up activity $i \in \mathcal{I}$, by using the same procedure that was used for solving subproblem (Q_t) in section 3.6. The reason for this is that $\Delta_i(\mathbf{k})$ might actually be smaller than $1/\gcd\{k_j \mid j \in J_i\}$, once the optimal values for k_j $(j \in \mathcal{J})$ are known. In this respect, it could also be interesting to repeat this procedure iteratively by defining a new approximate expression for $\Delta_i(\mathbf{k})$, which explicitly accounts for the discrepancy between the predicted and actual value. This procedure could be repeated until convergence in terms of a steady state or limit cycle is observed. We have chosen not to incorporate this iterative process, and to leave these suggestions for further research.

A dynamic programming formulation

In order to arrive at a dynamic programming formulation, let us denote with $h_i(k)$ the immediate costs associated with the assignment of set-up activity $i \in \mathcal{I}$ to maintenance period $k \in \mathbb{N}^*$, given that all lower-level set-up activities $i' \in S_i$ must be executed at integer multiples of k. Under the above-mentioned assumptions, this yields $\Delta_i(\mathbf{k}) = 1/k$, which on its turn leaves us with the following expression for $h_i(k)$:

$$h_i(k) = \frac{s_i \cdot t}{k} + \sum_{j \in J_i^*} \min_{l_j \geq 1} \{\Phi_j(k \cdot l_j / t)\}$$

Since $\Phi_j(x)$ is decreasing for $x < x_j^*$, and increasing for $x > x_j^*$, it is obvious that either $l_j^* = \lfloor x_j^* \cdot t/k \rfloor$ or $l_j^* = \lceil x_j^* \cdot t/k \rceil$ for all $j \in J_i^*$, and thus $h_i(k)$ can be determined rather straightforwardly. To continue our analysis, we denote with $g_i(k)$ the minimal costs of the subtree associated with set-up activity $i \in \mathcal{I}$, provided that it must be assigned to an integer multiple of maintenance period $k \in \mathbb{N}^*$, i.e. $\hat{k}_i \bmod k = 0$. Then it is easily verified that $g_i(k)$ can be determined recursively by means of the following dynamic programming formulation:

$$g_i(k) = \min_{l \geq 1} \left\{ h_i(k \cdot l) + \sum_{i' \in S_i^*} g_{i'}(k \cdot l) \right\}$$

The problem in the above formulation is that, in order to determine $g_i(k)$, an infinite number of alternative decisions $l \geq 1$ must be evaluated. In the remainder of this section, we will present an efficient procedure with which only a finite number of alternatives $l \in \{l_{\min}, ..., l_{\max}\}$ needs to be evaluated. This procedures starts with a promising, initial decision l^*. Subsequently, it proceeds iteratively by gradually increasing the set of possible decisions, until at some point we know for sure that all decisions $l < l_{\min}$ and $l > l_{\max}$ cannot be optimal. In the remainder of this section, this approach will be discussed in more detail.

Bounds for the optimal decision

As a starting point, we determine the optimal solution to the relaxed version of problem $g_i(k)$. Similar to (R), this yields a convex programming problem in a convex search space, which can easily be solved to optimality with the use of standard search techniques:

$$\check{g}_i(k) = \min_{\boldsymbol{\tau} \in \mathcal{T}: \tau_i \leq t/k} \sum_{i' \in \{i\} \cup S_i} \left\{ s_{i'} \cdot \tau_{i'} + \sum_{j \in J_i^*} \xi_j(\tau_{i'}) \right\}.$$

The underlying observation behind this formulation is that set-up activity $i \in \mathcal{I}$ must be assigned to an integer multiple of maintenance period $k \in \mathbb{N}^*$, and thus $\Delta_i(\mathbf{k}) \leq 1/k$ is known in advance. Based upon the optimal value of $\tau_i > 0$ in this relaxation, it is at least intuitively clear that $l^* = [t/(k \cdot \tau_i)]$ is one of the most promising decisions. Here, $[x] \in \mathbb{Z}$ denotes the nearest integer relative to $x \in \mathbb{R}$,

e.g. $[1.2] = 1$ and $[1.8] = 2$. Therefore, and in line with the above, an upper bound $\hat{g}_i(k \mid l = l^*)$ for $g_i(k)$ is determined with the use of dynamic programming:

$$\hat{g}_i(k \mid l) = h_i(k \cdot l) + \sum_{i' \in S_i^*} g_{i'}(k \cdot l)$$

Our analysis now proceeds by deriving a lower bound $\check{g}_i(k \mid l < l^*)$ and $\check{g}_i(k \mid l > l^*)$ for the remaining decisions $l < l^*$ and $l > l^*$, by using a formulation which is similar to $\check{g}_i(k)$. Here, we denote $\tau_{\min} = 1/(k \cdot (l^* - 1))$ and $\tau_{\max} = 1/(k \cdot (l^* + 1))$ for notational convenience:

$$\check{g}_i(k \mid l < l^*) = \min_{\boldsymbol{\tau} \in \mathcal{T}: \tau_i \geq \tau_{\min}} \sum_{i' \in \{i\} \cup S_i} \left\{ s_{i'} \cdot \tau_{i'} + \sum_{j \in J_i^*} \xi_j(\tau_{i'}) \right\}$$

$$\check{g}_i(k \mid l > l^*) = \min_{\boldsymbol{\tau} \in \mathcal{T}: \tau_i \leq \tau_{\max}} \sum_{i' \in \{i\} \cup S_i} \left\{ s_{i'} \cdot \tau_{i'} + \sum_{j \in J_i^*} \xi_j(\tau_{i'}) \right\}$$

The underlying observation behind these lower bounds is that we know for sure that $\Delta_i(\mathbf{k}) \geq 1/(k \cdot (l^* - 1))$ if $l < l^*$, and that $\Delta_i(\mathbf{k}) \leq 1/(k \cdot (l^* + 1))$ if $l > l^*$. Obviously, $\check{g}_i(k \mid l < l^*) \geq \hat{g}_i(k \mid l^*)$ implies that decisions $l < l^*$ can as well be neglected without affecting the problem. In a similar way, $\check{g}_i(k \mid l > l^*) \geq \hat{g}_i(k \mid l^*)$ implies that decisions $l > l^*$ can never lead to the optimal solution. In all other cases, further investigations into decisions $l < l^*$ and/or $l > l^*$are necessary. Under these circumstances, the most promising of $\hat{g}_i(k \mid l^* - 1)$ and $\hat{g}_i(k \mid l^* + 1)$, in view of the corresponding lower bounds $\check{g}_i(k \mid l < l^*)$ and $\check{g}_i(k \mid l > l^*)$, is determined with the use of dynamic programming.

After each iteration of this procedure, upper bounds $\hat{g}_i(k \mid l)$ have been derived for a finite set of alternative decisions $l \in \{l_{\min}, ..., l_{\max}\}$, where $1 \leq l_{\min} \leq l^* \leq l_{\max} < \infty$. Once again, it is now determined whether further investigation into decisions $l < l_{\min}$ or $l > l_{\max}$ is necessary. Since $\check{g}_i(k \mid l > l_{\max}) \rightarrow \infty$ as $l_{\max} \rightarrow \infty$, this procedure will terminate after a finite number of iterations.

An efficient procedure for $\check{g}_i(k)$, $\check{g}_i(k \mid l < l_{\min})$ and $\check{g}_i(k \mid l > l_{\max})$

The question remains how to (efficiently) determine $\check{g}_i(k)$, $\check{g}_i(k \mid l < l_{\min})$ and $\check{g}_i(k \mid l > l_{\max})$ during the course of our dynamic programming algorithm. It is easily verified from the underlying expressions, that it is sufficient to formulate an efficient procedure, with which problems $R_i(\alpha, \beta)$ of the following type can be solved to optimality:

$$R_i(\alpha,\beta) = \min_{\boldsymbol{\tau}\in\mathcal{T}:\alpha\leq\tau_i\leq\beta} \sum_{i'\in\{i\}\cup S_i} \left\{ s_{i'}\cdot\tau_{i'} + \sum_{j\in J_i^*} \xi_j(\tau_{i'}) \right\}$$

The advantage of this more general formulation is that $R_i(\alpha,\beta)$ has some nice structural properties, which are stated more explicitly in the following lemma:

Lemma 4 *$R_i(\tau,\tau)$ is convex on $\langle 0,\infty\rangle$, and has a unique minimum $0<\tau_i^*\leq\theta_i^*$.*

Proof. Following an analysis similar to Lemma 2, it is immediately clear that $R_i(\tau,\tau)$ has a unique minimum $0<\tau_i^*\leq\theta_i^*$. Moreover, $R_i(\tau,\tau)=s_i\cdot\tau+\sum_{j\in J_i^*}\xi_j(\tau)$ is convex on $\langle 0,\infty\rangle$ for each lowest-level set-up activity $i\in\mathcal{I}$ with $S_i^*=\varnothing$. The proof now proceeds with induction on $i\in\mathcal{I}$. To this end, we need the following recursive relationship for $R_i(\alpha,\beta)$, which more explicitly states that $\alpha\leq\tau_i\leq\beta$ and $\tau_{i'}\leq\tau_i$ for all $i'\in S_i$:

$$R_i(\alpha,\beta) = \min_{\alpha\leq\tau_i\leq\beta} \left\{ s_i\cdot\tau_i + \sum_{j\in J_i^*}\xi_j(\tau_i) + \sum_{i'\in S_i^*} R_{i'}(0,\tau_i) \right\}$$

Now consider an arbitrary set-up activity $i\in\mathcal{I}$, and suppose that $R_{i'}(\tau,\tau)$ is convex on $\langle 0,\infty\rangle$ for all lower-level set-up activities $i'\in S_i^*$. Then $R_{i'}(0,\tau)$ is convex on $\langle 0,\infty\rangle$ for all $i'\in S_i^*$, almost by definition, and thus $R_i(\tau,\tau)=s_i\cdot\tau+\sum_{j\in J_i^*}\xi_j(\tau)+\sum_{i'\in S_i^*}R_{i'}(0,\tau)$ is also convex on $\langle 0,\infty\rangle$. This completes the proof. □

Summarizing, $R_i(\alpha,\beta)$ and thus $\check{g}_i(k)=R_i(0,t/k)$, $\check{g}_i(k\mid l<l_{\min})=R_i(\tau_{\min},\infty)$ and $\check{g}_i(k\mid l>l_{\max})=R_i(0,\tau_{\max})$ can be determined recursively, provided that the optimal solutions $\tau_i^*>0$ to the optimization problems $(R_i)\sim R_i(0,\infty)$ are known in advance:

$$R_i(\alpha,\beta) = \begin{cases} s_i\cdot\alpha + \sum_{j\in J_i^*}\xi_j(\alpha) + \sum_{i'\in S_i^*} R_{i'}(0,\alpha) & \text{if } \tau_i^*\leq\alpha\leq\beta \\ s_i\cdot\tau_i^* + \sum_{j\in J_i^*}\xi_j(\tau_i^*) + \sum_{i'\in S_i^*} R_{i'}(0,\tau_i^*) & \text{if } \alpha\leq\tau_i^*\leq\beta \\ s_i\cdot\beta + \sum_{j\in J_i^*}\xi_j(\beta) + \sum_{i'\in S_i^*} R_{i'}(0,\beta) & \text{if } \alpha\leq\beta\leq\tau_i^* \end{cases}$$

From this recursive formulation, it is immediately clear that only $|S_i|+1$ multiplications and $|J_i|$ function evaluations are required to determine $R_i(\alpha,\beta)$, for arbitrary values of α and β. Of course, this is an extremely useful property from a computational point of view. The question remains how to determine $\tau_i^*>0$ for each set-up

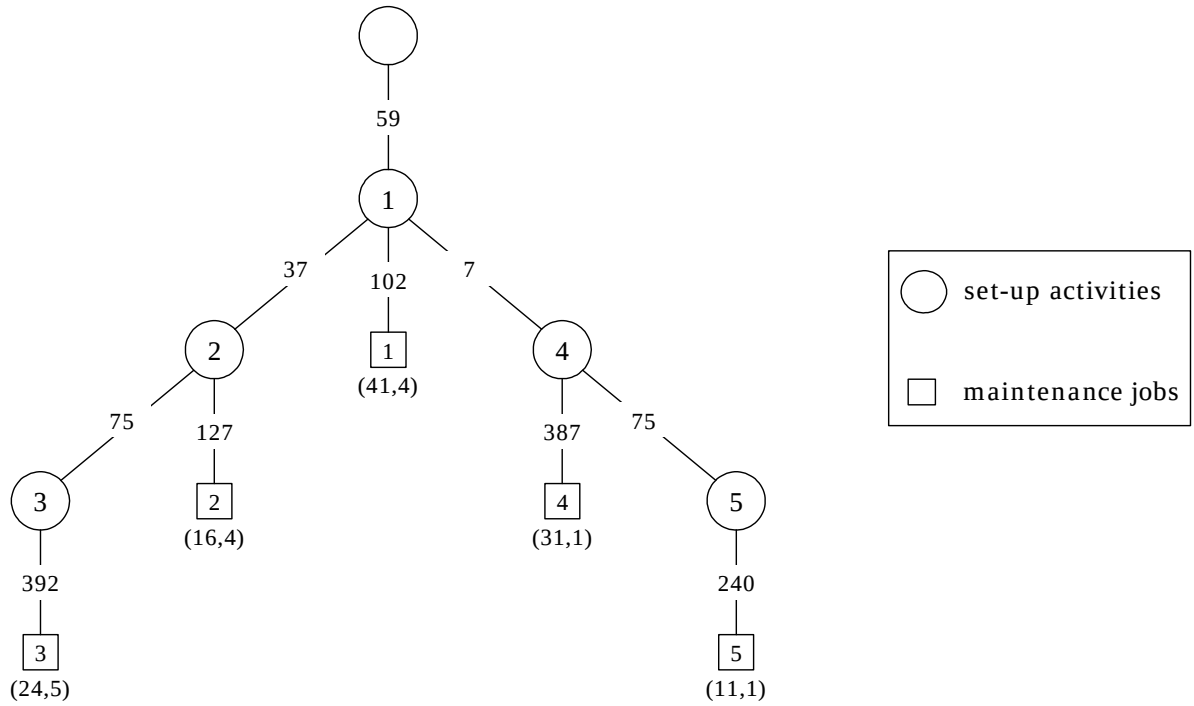


Figure 3.3: Example of a test problem consisting of 5 set-up activities and 5 components. Set-up and maintenance costs s_i and c_j are shown at the arcs, deterioration cost parameters a_j and b_j in brackets at the corresponding nodes.

activity $i \in \mathcal{I}$. Similar to $(R) \equiv (R_1)$, each (R_i) is a convex programming problem in a convex search space, which can easily be solved to optimality with the use of standard search techniques. Given the optimal solution $(t_1^*, ..., t_m^*)$ to problem (R), it is now immediately clear that $t_i^* > t_{i'}^*$ for some $i \in \mathcal{I}$ and $i' \in S_i^*$ implies that $\tau_{i'}^* = t_{i'}^*$. In a similar way, mutual relationships can be derived for the optimal solutions to problems (R_i) and $(R_{i'})$, where $i \in \mathcal{I}$ and $i' \in S_i^*$. Further details are skipped, since they are not so relevant for what follows.

Characteristics of the optimal decision

Of course, it depends on the tightness of these lower bounds, as well as the characteristics of the problem under consideration, how much alternative decisions must be evaluated in each decision problem $g_i(k)$. Nevertheless, computational results within a variety of (large) test problems (see section 3.9) showed that only one decision was needed in approximately 97.8% of all test problems. In the remaining 2.2% of all other test problems, exactly two decisions were required. Finally, three alternative decisions had to be evaluated in approximately 0.01% of all cases.

Table 3.3: Parameter settings for the numerical example of Figure 3.3.

i, j	s_i	a_j	b_j	c_j	x_j^*	y_j^*	θ_i^*	t_i^*	τ_i^*
1	59	41	4	102	0.909	0.996	1.100	1.004	1.004
2	37	16	4	127	1.147	1.284	0.872	0.828	0.828
3	75	24	5	392	1.218	1.294	0.821	0.797	0.797
4	7	31	1	387	3.533	3.823	0.283	0.280	0.280
5	75	11	1	240	4.671	5.885	0.214	0.187	0.187

We conclude this section by observing that the optimal decision to subproblem $g_i(k)$ often contains useful information for other subproblems $g_i(k')$, where $k' > k$. More specifically, if we denote with $l_i^*(k) \geq 1$ the optimal decision for subproblem $g_i(k)$, it is immediately clear that $l_i^*(k) = p \cdot q \geq 1$ for some $p, q \geq 1$ implies that $l_i^*(p \cdot k) = q$ and $l_i^*(q \cdot k) = p$. After all, if the optimal maintenance period equals 4 provided that it must be an integer multiple of 2, then it also equals 4 if it must be an integer multiple of 4. Of course, this nice structural property was further exploited during the implementation of our dynamic programming algorithm.

3.8 Numerical example

Consider a production system consisting of $m = 5$ set-up activities and $n = 5$ components, as shown in Figure 3.3. For each component $j \in \mathcal{J}$, we consider a deterioration cost function $M_j(\cdot)$ of the following form $M_j(x) = a_j \cdot x^{b_j}$, where $a_j > 0$ and $b_j > 0$ are strictly positive constants. The costs s_i of set-up activities $i \in \mathcal{I}$, as well as the parameters (a_j, b_j, c_j) for components $j \in \mathcal{J}$, are depicted in Figure 3.3. Moreover, the corresponding values of x_j^* and y_j^* for each component $j \in \mathcal{J}$, as well as the values of θ_i^*, t_i^* and τ_i^* for each set-up activities $i \in \mathcal{I}$, are summarized in Table 3.3.

As a starting point, our iterative heuristic determines the optimal solution to problem (R). This yields $t_1^* = 1.004$, $t_2^* = 0.828$, $t_3^* = 0.797$, $t_4^* = 0.280$, and $t_5^* = 0.187$, with corresponding lower bound 1157.42. Subsequently, we initialize $t = 1.004$, and subproblems (Q_k) and (Q_t) are solved repeatedly, until no improvements are observed in two consecutive iterations. The results of this iterative process for both heuristics are depicted in Table 3.4.

Clearly, the finite-set heuristic generates a maintenance cycle of $\text{lcm}\{1, 3, 5\}/t = 15/0.859 \approx 17.46$ time units. Within this maintenance cycle, total maintenance costs amount to $17.46 \cdot 1176.36 \approx 20542$ on average, whereas preventive set-up costs equal $15 \cdot \{59 + 37 + 75 + \frac{7}{15} \cdot 7 + \frac{1}{5} \cdot 75\} = 2839$, i.e. approximately 13.8% of total

Table 3.4: Consecutive iterations of the finite-set heuristic (FSH), and the structured-set heuristic (SSH), for the numerical example of Figure 3.3.

		t	$\{k_1, ..., k_5\}$	$\{\Delta_1, ..., \Delta_5\}$	costs
	(Q_k)	1.004	$\{1,1,1,4,6\}$	$\{1,1,1,\frac{1}{3},\frac{1}{6}\}$	1216.10
	(Q_t)	0.878	$\{1,1,1,4,6\}$	$\{1,1,1,\frac{1}{3},\frac{1}{6}\}$	1186.78
(FSH)	(Q_k)	0.878	$\{1,1,1,3,5\}$	$\{1,1,1,\frac{7}{15},\frac{1}{5}\}$	1177.34
	(Q_t)	0.859	$\{1,1,1,3,5\}$	$\{1,1,1,\frac{7}{15},\frac{1}{5}\}$	1176.36
	(Q_k)	0.859	$\{1,1,1,3,5\}$	$\{1,1,1,\frac{7}{15},\frac{1}{5}\}$	1176.36
	(Q_k)	1.004	$\{1,1,1,4,6\}$	$\{1,1,1,\frac{1}{3},\frac{1}{6}\}$	1217.27
	(Q_t)	0.878	$\{1,1,1,4,6\}$	$\{1,1,1,\frac{1}{3},\frac{1}{6}\}$	1187.81
(SSH)	(Q_k)	0.878	$\{1,1,1,3,6\}$	$\{1,1,1,\frac{1}{3},\frac{1}{6}\}$	1179.80
	(Q_t)	0.864	$\{1,1,1,3,6\}$	$\{1,1,1,\frac{1}{3},\frac{1}{6}\}$	1179.28
	(Q_k)	0.864	$\{1,1,1,3,6\}$	$\{1,1,1,\frac{1}{3},\frac{1}{6}\}$	1179.28

maintenance costs. Similarly, the structured-set heuristic generates a maintenance cycle of $\text{lcm}\{1,3,6\}/t = 6/0.864 \approx 6.94$ time units. Within this maintenance cycle, total maintenance costs amount to $6.94 \cdot 804.29 \approx 8189$ on average, whereas preventive set-up costs equal $6 \cdot \{59 + 37 + 75 + \frac{1}{3} \cdot 7 + \frac{1}{6} \cdot 75\} = 1115$, i.e. approximately 13.6% of total maintenance costs. Moreover, the guaranteed performance of the finite-set and structured-set heuristic, i.e. the maximal deviation with respect to the lower bound, equals only 1.64% and 1.89% respectively.

3.9 Computational results

In this section, we will discuss the results of a series of numerical experiments, that were carried out to investigate the performance of both heuristics. To be specific, we tested our heuristics on several test problems, in which the number of set-up activities, the number of components, and the corresponding costs were varied randomly. In order to avoid that $\Delta_i(\mathbf{k}) = 1$ for all $i \in \mathcal{I}$ as much as possible, we chose to attach exactly one component to each set-up activity, i.e. $|J_i^*| = 1$ for all $i \in \mathcal{I}$. As a consequence, the number of set-up activities equals the number of components in each test problem. Moreover, set-up activities were attached to each other by choosing the parent of set-up activity $i \in \mathcal{I}$ randomly among set-up activities $\{1, ..., i-1\}$. It is easily observed that this procedure allows for each possible set-up structure that can be thought of.

Table 3.5: Computational results for the finite-set heuristic (FSH) and the structured-set heuristic (SSH), based on 1000 randomly generated test problems with $m = n = 5$, $s_i \in (1, 100)$, $a_j \in (100, 500)$, $b_j \in (10, 50)$ and $c_j \in (1, 5)$.

	finite-set heuristic			structured-set heuristic		
	minimal	average	maximal	minimal	average	maximal
# iterations	3	3.31	7	3	3.31	7
CPU time (s)	0	2.95	16	0	0.01	1
performance (%)	0.01	2.03	9.56	0.01	2.04	9.56
interval length	0.83	1.24	3.14	0.83	1.24	3.14
cycle length	1.09	5.03	78.18	1.09	5.02	78.18

3.9.1 Small test problems

First of all, we investigated the performance of both heuristics for a series of small test problems, that could be solved by both heuristics within reasonable computation times. Obviously, this required the number of maintenance opportunities $|\mathcal{L}^*|$ in the mixed integer linear programming formulation not to become too large. In each test problem, this was achieved by defining $m = 5$ set-up activities and $n = 5$ components, and by choosing s_i, a_j, b_j and c_j in such a way that the corresponding values of x_j^* and y_j^* were not subject to extreme fluctuations. To be specific, the costs of set-up activities $i \in \mathcal{I}$ were drawn at random from $s_i \in (1, 100)$, whereas the costs of component $j \in \mathcal{J}$ were drawn at random from $a_j \in (100, 500)$, $b_j \in (10, 50)$, and $c_j \in (1, 5)$. Under these conditions, it could be shown that $0.85 \leq x_j^* \leq 7.08$ and $0.85 \leq y_j^* \leq 10$ for all $j \in \mathcal{J}$.

In each test problem, we administrated the computation time and number of iterations needed by the finite-set and structured-set heuristic. Moreover, we calculated the relative performance of both heuristics in terms of the deviation with respect to the lower bound. Finally, we took a closer look at the solutions found, in terms of the interval length $1/t$, and cycle length $\text{lcm}(k_1, ..., k_n)/t$. From the results in Table 3.5, we conclude that the differences between the finite-set and structured-set heuristic are in general very small. In fact, a closer look at the results showed that both heuristics generated identical maintenance cycles in 988 out of 1000 test problems. In the remaining 12 test problems (see 3.6), the finite-set heuristic always outperformed the structured-set heuristic, with a maximum of 0.70%. With respect to the (guaranteed) performance of both heuristics, we claim that an average deviation of 2.04% from the lower bound is quite satisfactory. After all, one must realize that this deviation also includes the gap between the lower bound and the optimal solution.

Table 3.6: Computational results for 12 out of 1000 test problems in which the finite-set and structured-set heuristic generated different solutions.

finite-set heuristic			structured-set heuristic		
t^*	$\{k_1^*, ..., k_5^*\}$	$\{\Delta_1^*, ..., \Delta_5^*\}$	t^*	$\{k_1^*, ..., k_5^*\}$	$\{\Delta_1^*, ..., \Delta_5^*\}$
0.94	$\{1, 4, 3, 2, 1\}$	$\{1, \frac{1}{4}, \frac{2}{3}, \frac{1}{2}, 1\}$	0.95	$\{1, 4, 4, 2, 1\}$	$\{1, \frac{1}{4}, \frac{1}{2}, \frac{1}{2}, 1\}$
1.04	$\{1, 1, 4, 2, 3\}$	$\{1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}\}$	0.91	$\{1, 1, 3, 1, 2\}$	$\{1, 1, 1, 1, \frac{1}{2}\}$
1.13	$\{1, 1, 2, 3, 2\}$	$\{1, 1, \frac{1}{2}, \frac{2}{3}, \frac{1}{2}\}$	1.11	$\{1, 1, 2, 2, 2\}$	$\{1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\}$
0.89	$\{1, 2, 1, 1, 3\}$	$\{1, \frac{2}{3}, 1, 1, \frac{1}{3}\}$	0.90	$\{1, 2, 1, 1, 4\}$	$\{1, \frac{1}{2}, 1, 1, \frac{1}{4}\}$
0.99	$\{1, 3, 2, 1, 1\}$	$\{1, \frac{2}{3}, \frac{1}{2}, 1, 1\}$	0.97	$\{1, 2, 2, 1, 1\}$	$\{1, \frac{1}{2}, \frac{1}{2}, 1, 1\}$
0.71	$\{1, 2, 1, 3, 1\}$	$\{1, \frac{2}{3}, 1, \frac{1}{3}, 1\}$	0.74	$\{1, 3, 1, 3, 1\}$	$\{1, \frac{1}{3}, 1, \frac{1}{3}, 1\}$
0.85	$\{1, 1, 3, 2, 2\}$	$\{1, 1, \frac{2}{3}, \frac{1}{2}, \frac{1}{2}\}$	0.93	$\{1, 1, 2, 2, 2\}$	$\{1, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\}$
0.86	$\{1, 1, 1, 3, 5\}$	$\{1, 1, 1, \frac{7}{15}, \frac{1}{5}\}$	0.86	$\{1, 1, 1, 3, 6\}$	$\{1, 1, 1, \frac{1}{3}, \frac{1}{6}\}$
0.76	$\{1, 1, 2, 3, 3\}$	$\{1, 1, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}\}$	0.74	$\{1, 1, 2, 3, 2\}$	$\{1, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{2}\}$
0.91	$\{1, 1, 1, 3, 2\}$	$\{1, 1, 1, \frac{2}{3}, \frac{1}{2}\}$	0.89	$\{1, 1, 1, 2, 2\}$	$\{1, 1, 1, \frac{1}{2}, \frac{1}{2}\}$
0.86	$\{1, 3, 1, 2, 1\}$	$\{1, \frac{2}{3}, 1, \frac{1}{2}, 1\}$	0.87	$\{1, 4, 1, 2, 1\}$	$\{1, \frac{1}{2}, 1, \frac{1}{2}, 1\}$
0.87	$\{5, 3, 2, 1, 3\}$	$\{1, 1, \frac{2}{3}, 1, \frac{1}{3}\}$	0.89	$\{5, 3, 2, 1, 4\}$	$\{1, 1, \frac{1}{2}, 1, \frac{1}{4}\}$

3.9.2 Large test problems

As mentioned before, the finite-set heuristic (FSH) becomes inattractive, or even intractable, if the set of possible maintenance periods becomes too large, e.g. $\mathcal{K} = \{1, ..., 100\}$. In that case, we have to use the structured-set heuristic (SSH), in order to obtain reasonable solutions within acceptable computation times. To this end, we carried out another series of 1000 test problems, but this time with $m = 25$ set-up activities and $n = 25$ components. In each test problem, the costs of set-up activities $i \in \mathcal{I}$ were drawn at random from $s_i \in (1, 1000)$. Moreover, the costs of component $j \in \mathcal{J}$ were chosen randomly from $a_j \in (1, 1000)$, $b_j \in (1, 100)$, and $c_j \in (1, 10)$. Under these conditions, it could be shown that $0.10 \leq x_j^* \leq 31.62$ and $0.14 \leq y_j^* \leq 161.25$ for all $j \in \mathcal{J}$, which clearly complicates the problem.

From the results in Table 3.7, it can be observed that the computation time and number of iterations needed by the structured-set heuristic have increased. Nevertheless, they did not grow explosively with the size and complexity of the test problems under consideration. This is a potentially valuable insight, since the latter implies that large test problems can be tackled by our dynamic programming algorithm within reasonable computation times. On the other hand, the length of the corresponding maintenance cycle increased significantly, whereas the performance decreased slightly. Although these observations were to be expected, we claim that

Table 3.7: Computational results for the structured-set heuristic (SSH), based on 1000 randomly generated test problems with $m = n = 25$, $s_i \in (1, 1000)$, $a_j \in (1, 1000)$, $b_j \in (1, 100)$ and $c_j \in (1, 10)$.

	minimal	average	maximal
# iterations	3	4.96	13
CPU time (s)	0	0.30	1
performance (%)	0.68	3.39	18.32
interval length	0.80	1.13	1.26
cycle length	1.16	47.59	1917.58

an average deviation of 3.39% from the lower bound is also quite satisfactory for practical purposes. Once again, a significant part of this deviation may be due to the deviation between the lower bound an the optimal solution.

3.10 Concluding remarks

In this chapter, we presented a variety of optimization techniques which can assist maintenance planners in the design of (near-)optimal preventive maintenance cycles, for a multi-component production system with multiple interrelated set-up activities. To this end, a powerful modelling framework was presented, in which each component is maintained preventively at integer multiples of a certain basis maintenance interval, which is the same for all components. For a given basis maintenance interval, we developed a mixed integer linear programming as well as a dynamic programming formulation, which can be used to determine an (optimal) maintenance period for each component.

Subsequently, these methods were incorporated in an iterative heuristic approach, which can be applied if the basis maintenance interval is also free to choose. Based on a series of 1000 randomly generated (small) test problems, we concluded that the difference between both methods is in general very small, and that the dynamic programming formulation should be preferred for its efficiency. Besides, another series of (large) test problems pointed out that this approach generates near-optimal solutions within reasonable computation times, even for production systems with a large number of set-up activities and/or components.

Summarizing, we claim that the problems addressed in this chapter have been solved satisfactorily by our methods. It is not difficult, however, to come up with a number of promising extensions to our modelling framework. As a starting point,

the model discussed here might be classified as an additive one, i.e. the costs of maintaining a certain group of components are modelled as the sum of the individual costs for each component. There are situations, however, where time is the most crucial and expensive factor. In such cases, it would also make sense to model the time required for a certain group of components as the maximum of the individual times required for each component. Of course, it would be interesting to develop a good working algorithm for such cases.

Within our modelling framework, we assumed - without loss of generality - that each component is maintained preventively at the end of each maintenance cycle. As a consequence, the peak workload associated with this maintenance cycle may be significantly larger than is strictly necessary. In this respect, it might also be worthwile to investigate the possibilities for adjusting the maintenance cycle, such that the peak workload is minimized, but overall maintenance costs are not affected. For similar reasons, there is a potential of challenging optimization problems if the maintenance cycles of different production systems are combined into an overall maintenance cycle, in which the peak workload for the maintenance department is minimized.

Finally, it would be interesting to include the possibility of frequency-constrained maintenance jobs, or equivalently components that must be maintained preventively at prescribed or smaller intervals (see previous chapter). Within our modelling framework, these frequency constraints could be incorporated implicitly, by adjusting the individual deterioration cost function of these components. Moreover, they could be modelled explicitly, by defining a set of feasible maintenance intervals and/or maintenance periods within each of the algorithms discussed in this chapter.

Chapter 4

Preventive maintenance and the interval availability distribution of an unreliable production system

Traditionally, the optimal preventive maintenance interval for an unreliable production system has been determined by maximizing its limiting availability. Nowadays, it is widely recognized that this performance measure does not always provide relevant information for practical purposes, and the so-called interval availability distribution is often seen as a more appropiate performance measure. Surprisingly enough, the relation between preventive maintenance and interval availability has received little attention in existing literature. In this paper, a series of mathematical models and optimization techniques is presented, with which the optimal preventive maintenance interval can be determined from an interval availability point of view, rather than from a limiting availability perspective. Computational results for a class of representative test problems indicate that significant improvements of up to 30% in the guaranteed interval availability can be obtained, by increasing preventive maintenance frequencies somewhere between 10% and 70%.

4.1 Introduction

In studying the performance of an unreliable production system, the limiting availability does not always provide the most relevant information for practical purposes. For example, the amount of gas to be delivered over a finite period of time is often contractually guaranteed in the oil industry (Aven 1993). Although short interruptions of the production process can usually be covered by inventory backups, a loss

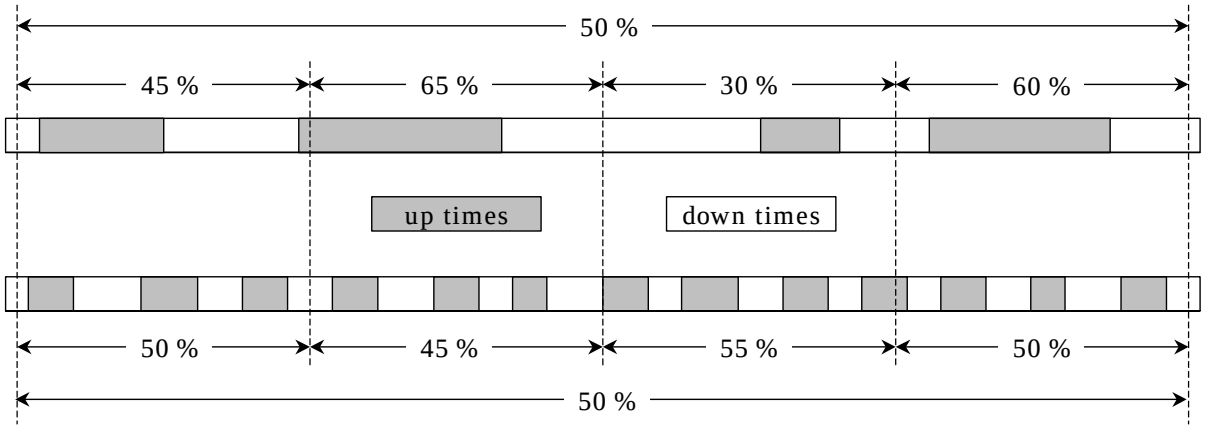


Figure 4.1: Production systems with similar limiting availabilities, but different interval availability distributions.

of production for several consecutive days might cause problems in meeting the sales contract, involve high penalty costs, and - in the worst case - loss of goodwill or even customers (Van Rijn and Schornagel 1987). In computer and manufacturing systems, the guaranteed performance during a finite period of time is sometimes a more important and even competitive factor, than the average performance observed over an infinite horizon (Goyal and Tantawi 1988). In this respect, the **interval availability** of a production system is often seen as a more appropiate performance measure in a practical context. This is particularly true for order-driven manufacturing systems, in which capacity planning plays a key strategical role in satisfying contractual obligations.

Most capacity planning tools used in industry account for random outages by computing average capacity in terms of limiting availability. By doing so, it is immediately clear that during a given period of time (e.g. a week), capacity problems will occur in approximately 50% of all cases. Since this is generally not acceptable, a safety margin is usually build in, in order to ensure satisfactory capacity in e.g. at least 95% of all cases. However, even if this works well in practice, it underlines the point that thinking in terms of the guaranteed capacity of a production system during a finite period of time, is often more appropriate than thinking about its average capacity in the long run. In this respect, a production system with frequent, predictable and short interruptions is to be preferred above one with infrequent, unpredictable and long interruptions, all other things being equal (see Figure 4.1). This is a potentially valuable insight, since random breakdowns are one of the major sources of variability.

During the last decades, this and other factors have resulted in an increased popularity of mathematical models for reliability and maintenance optimization, e.g. see McCall (1965), Pierskalla and Voelker (1976), Sherif and Smith (1981) , Valdez-Flores and Feldman (1989), Cho and Parlar (1991) and Dekker et al. (1997) for extensive literature reviews. At the same time, a growing interest could be observed in modelling the short term behavior of production systems, in terms of the so-called interval availability distribution. The reader is referred to Smith (1997b) for a comprehensive and up-to-date survey on existing literature. Surprisingly enough, the interactions between preventive maintenance on the one hand, and interval availability on the other hand, have received little attention in existing literature, possibly because of the inherent mathematical complications.

If a production system is repaired at failure, and thus all maintenance is corrective, consecutive up (life) and down (repair) times are usually modelled as stochastically independent random variables. Obviously, this modelling assumption cannot be sustained if preventive maintenance is carried out at regular intervals. In that case, consecutive up and down times become mutually dependent random variables, since small up times (due to failures) are usually followed by large down times (due to repairs), and vice versa. Obviously, this phenomenon does not make life easier from a mathematical point of view. But even for a two-state production system without preventive maintenance, i.e. with alternating and mutually independent up and down times, closed-form solutions for the interval availability distribution are not available. Pioneering work on this subject was carried out by Takács (1957), who derived an analytical expression consisting of an infinite summations of terms, each consisting of multiple convolutions of the life and repair time distributions. Since then, several authors have tried to find reasonable approximations, as well as lower and upper bounds for the interval availability distribution, e.g. see De Souza e Silva and Gail (1986), Van der Heijden (1987), Van Rijn and Schornagel (1987), Van der Heijden and Schornagel (1988), De Souza e Silva and Gail (1989), Wartenhorst (1993), Csenki (1995), Haukaas and Aven (1996) and Smith (1997a).

Up to our knowledge, the derivation of analytical expressions for the interval availability distribution of a two-state production system, which is maintained preventively at regular intervals according to an age replacement strategy, has not been subject of any study in existing literature. Only Schäbe (1996) considers a somewhat similar problem, in which the time between two consecutive preventive maintenance actions is modelled as a random variable with known distribution function. In general, this yields a much simpler modelling framework, since the initiation of preventive maintenance does not depend on the state of the underlying production system.

4.2 General approach

Consider an unreliable production system which is repaired upon failure, and maintained preventively as soon as $\theta > 0$ time units have elapsed since the last maintenance action, either preventive or corrective. After preventive and/or corrective maintenance, the system can be considered as good as new. The time to failure or lifetime L of the production system is described by a cumulative distribution function $F(.)$, with probability density function $f(.)$, and corresponding mean $\mu_L > 0$ and variance $\sigma_L^2 \geq 0$. Moreover, the preventive maintenance time P is described by a cumulative distribution function $G(.)$, with mean $\mu_P > 0$ and variance $\sigma_P^2 \geq 0$. Finally, the corrective maintenance (repair) time R is described by a cumulative distribution function $H(.)$, with mean $\mu_R > 0$ and variance $\sigma_R^2 \geq 0$. As in most maintenance optimization models, we assume that both L, P and R are mutually independent random variables.

4.2.1 Limiting availability

The limiting availability A_∞ is defined as the fraction of time that the production system is operational (up), if observed over an infinite period of time. If we denote with T_{up} a continuous period of time during which the system is operational, and with T_{down} a continuous period of time during which the system is not operational, then it follows from renewal theory (Cox 1962) that the limiting availability A_∞ is determined as:

$$A_\infty = \frac{E\{T_{up}\}}{E\{T_{up}\} + E\{T_{down}\}}$$

Depending on the length of the preventive maintenance interval $\theta > 0$, the following expressions can be derived for $E\{T_{up}\}$ and $E\{T_{down}\}$. Here, we denote $\overline{F}(\theta) = 1 - F(\theta)$ for notational convencience:

$$E\{T_{up}\} = \theta \cdot \overline{F}(\theta) + \int_0^\theta \tau \cdot f(\tau)\, d\tau$$

$$E\{T_{down}\} = \mu_P \cdot \overline{F}(\theta) + \mu_R \cdot F(\theta)$$

If $\theta \to \infty$, this yields $E\{T_{up}\} = \mu_L$, $E\{T_{down}\} = \mu_R$, and thus $A_\infty = \mu_L/(\mu_L + \mu_R)$. Traditionally, the optimal preventive maintenance for a production system has been determined in view of maximizing its limiting availiability. Unfortunately, this

performance measure does not always provide sufficient and relevant information for practical purposes. Sometimes, the so-called interval availability is seen as a more appropiate performance measure.

4.2.2 Interval availability

The interval availability is defined as the fraction of time that a production system is operational during a given time interval of finite length. Of course, it depends on the initial state of the system at the beginning of this interval, which type of behavior will be observed. From now on, we will assume - without loss of generality - that the production system starts as new at time $t = 0$. If we denote with U_t the cumulative up time during the interval $[0, t]$, then the interval availability A_t during this interval is defined as follows:

$$A_t = \frac{U_t}{t}$$

With $T_u = \inf\{t \mid U_t \geq u\}$, we denote the time required to attain a cumulative up time of u time units. Since both U_t and T_u are random variables, we are mainly interested in their cumulative distribution functions $P(U_t \leq u)$ and $P(T_u \leq t)$. By observing that $P(U_t \geq u) = P(T_u \leq t)$, it is sufficient to determine either $P(U_t \leq u)$ or $P(T_u \leq t)$. Up to our knowledge, and for no specific reason, mathematical models for the interval availability distribution have always been formulated in terms of U_t rather than T_u. From a theoretical point of view, however, the cumulative distribution function $P(T_u \leq t)$ is to be preferred, since the corresponding analytical expressions are mathematically more tractable (see the next section).

To avoid confusion, we will refer to $P(U_t \leq u)$ as the interval availability distribution, and to $P(T_u \leq t)$ as the **availability interval distribution**. Simply stated, $P(T_u \leq t)$ reflects the probability of completing a cumulative workload of u time units within t units of calendar time. Nowadays, this performance measure could be of considerable interest in e.g. due date determination and order acceptation, since it may provide useful information about the probability that a certain amount of workload will be completed within a certain amount of time.

As an illustrative example, consider a customer order of 10 hours processing time which must be completed within 3 days. Moreover, suppose that already 50 hours of workload have been accepted for other customers, with their own due dates as well. In case of a first-in-first-out (FIFO) service discipline, which is completely natural in such a setting, this would imply that on-time delivery of this new customer order can be realized with probability $P(T_{50+10} \leq 3 \cdot 24) = P(T_{60} \leq 72)$. Of course, it is up

to management to decide whether or not this is acceptable. Nevertheless, our model could provide useful decision support in this respect. Moreover, it could also be used to explore the the opportunities for, and consequences of changing priorities between customer orders.

4.2.3 Outline

As a starting point, we investigate the initial behavior of the system in section 4.3. To be specific, an analytical expression is derived for the probability $P_0(T_u \leq t)$ of at least u units of cumulative up time during the interval $[0, t]$. Subsequently, we investigate the limiting behavior of the production system in section 4.4, by deriving an analytical expression for the probability $P_\infty(T_u \leq t)$ of at least u units of cumulative up time during an arbitrary interval of length $t > 0$ in a stabilized situation. In section 4.5, some explicit formulas are derived for a production system with Gamma distributed repair and fixed maintenance times. Moreover, a simple but efficient algorithm is presented with which the optimal maintenance interval can be determined to a sufficient level of detail. Subsequently, a series of numerical experiments is presented in sections 4.6 and 4.7. Computational results indicate that significant improvements can be obtained in practice, if the optimal preventive maintenance interval is determined from an interval availability rather than a limiting availability point of view. Finally, section 4.8 summarizes some conclusions, and identifies some opportunities for further research.

4.3 Initial behavior of the system

As a starting point, we consider the case where the production system starts with an up time at time $t = 0$, and preventive maintenance is carried out as soon as the system has been operational (up) for $\theta > 0$ time units. In this respect, a clear distinction must be made between the case $u > \theta$, and the case $u \leq \theta$, since the latter requires much simpler modelling techniques.

4.3.1 Model without preventive maintenance ($u \leq \theta$)

If the required cumulative up time $u \leq \theta$, it is immediately clear that no preventive maintenance actions will be involved. As a consequence, all maintenance (if any) will be corrective, and our analysis becomes similar to the well-known failure-based model (Takács 1957). Since life times L and repair times R are stochastically inde-

pendent random variables, with corresponding cumulative distribution functions $F(.)$ and $H(.)$, this yields the following expression for $P_0(T_u \leq t)$:

$$P_0(T_u \leq t) = \sum_{n=0}^{\infty} H_n(t-u) \cdot \{F_n(u) - F_{n+1}(u)\}$$

Here, $F_n(.)$ and $H_n(.)$ denote the n-fold Stieltjes convolutions of $F(.)$ and $H(.)$ respectively, i.e. $F_1(u) = F(u)$ and $F_n(u) = \int_0^u f(v) \cdot F_{n-1}(u-v)\, dv$. More specifically, $F_n(u) - F_{n+1}(u)$ denotes the probability of exactly n failures during the first u units of cumulative up time, whereas the accumulated down time of these failures does not exceed the amount of $t-u$ time units with probability $H_n(t-u)$. Obviously, this analysis cannot be sustained as soon as subsequent up and down times become mutually dependent random variables. Of course, this happens if preventive maintenance is carried out at regular intervals. In that case, the correlation between consecutive up and down times usually drops below zero, since typically small (corrective) up times go together with large (corrective) down times, and large (preventive) up times go together with small (preventive) down times. Here, the correlation $\rho(T_{up}, T_{down})$ between consecutive up and down times is defined as follows:

$$\rho(T_{up}, T_{down}) = \frac{E\{T_{up} \cdot T_{down}\} - E\{T_{up}\} \cdot E\{T_{down}\}}{\sqrt{E\{T_{up}^2\} - E\{T_{up}\}^2} \cdot \sqrt{E\{T_{down}^2\} - E\{T_{down}\}^2}}$$

With $E\{T_{up}\}$ and $E\{T_{down}\}$ as defined earlier, $\rho\{T_{up}, T_{down}\}$ can be determined straightforwardly by observing that analytical expressions for $E\{T_{up} \cdot T_{down}\}$, $E\{T_{up}^2\}$, and $E\{T_{down}^2\}$ are also readily available:

$$E\{T_{up} \cdot T_{down}\} = \theta \cdot \mu_P \cdot \overline{F}(\theta) + \int_0^{\theta} \tau \cdot \mu_R \cdot f(\tau)\, d\tau$$

$$E\{T_{up}^2\} = \theta^2 \cdot \overline{F}(\theta) + \int_0^{\theta} \tau^2 \cdot f(\tau)\, d\tau$$

$$E\{T_{down}^2\} = (\mu_P^2 + \sigma_P^2) \cdot \overline{F}(\theta) + (\mu_R^2 + \sigma_R^2) \cdot F(\theta)$$

An illustrative example for this phenomenon is presented in Figure 4.2, which depicts the correlation between consecutive up and down times in a production system with Gamma distributed life times ($\mu_L = 40$, $\sigma_L = 10$), repair times ($\mu_R = 15$,

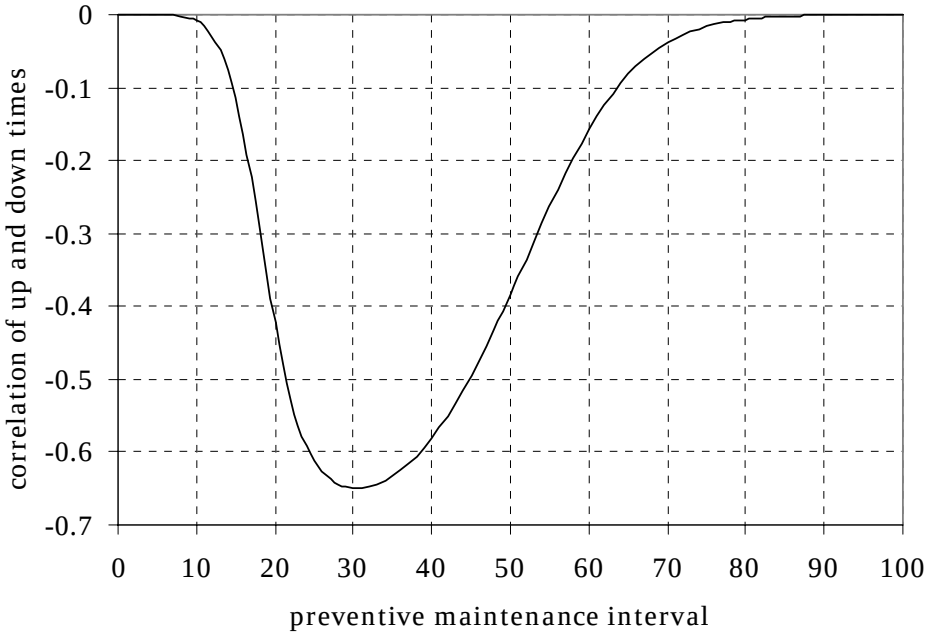


Figure 4.2: Correlation between consecutive up and down times in a production system with Gamma-distributed life, repair, and maintenance times, as a function of the preventive maintenance interval (example with $\mu_L = 40$, $\sigma_L = 10$, $\mu_P = 5$, $\sigma_P = 1$, $\mu_R = 15$ and $\sigma_R = 5$).

$\sigma_L = 5$), and maintenance times ($\mu_P = 5$, $\sigma_P = 1$), for a variety of preventive maintenance intervals. Apparently, the correlation between consecutive up and down times reduces to zero as the preventive maintenance interval goes to zero or infinity. This is intuitively clear, since a fully preventive or corrective maintenance strategy should lead to independent up (life) and down (maintenance or repair) times. On the other hand, if the preventive maintenance interval is close to the expected life time of the production system, the correlation between consecutive up and down times becomes relatively large. Simply stated, this phenomenon is caused by the fact that in this range, the uncertainty with respect to the occurence of either preventive or corrective maintenance actions, attains its highest possible level.

4.3.2 Model with preventive maintenance ($u > \theta$)

If the required cumulative up time satisfies $u > \theta$, our analysis proceeds as follows. First of all, we determine the probability $\xi_u^0(m, n)$ of exactly m preventive maintenance actions and n corrective maintenance actions during the first u units

of cumulative up time. Obviously, not all values of m and n correspond to non-zero probabilities $\xi_u^0(m,n)$. As a starting point, since each preventive maintenance action corresponds to exactly θ units of up time, the number of preventive maintenance actions m should at least satisfy $m \cdot \theta < u$. Moreover, since each corrective maintenance action corresponds to at most θ units of up time, the number of preventive and corrective maintenance actions $m+n$ should also satisfy $(m+n+1) \cdot \theta \geq u$.

In all other cases, it is possible to derive an analytical expression for $\xi_u^0(m,n)$. By the complete randomness of consecutive maintenance actions, i.e. preventive with probability $\overline{F}(\theta)$ and corrective with probability $F(\theta)$, the probability $\xi_u^0(m,n)$ must be equal to $\binom{m+n}{m}$ times the probability that exactly m consecutive preventive maintenance actions are followed by exactly n consecutive corrective maintenance actions within the first u units of cumulative up time. If we denote with $\tilde{F}(t) = P(L \leq t \mid L \leq \theta) = \min\{1, F(t)/F(\theta)\}$ the conditional cumulative lifetime distribution function, this yields the following expression for $\xi_u^0(m,n)$. Here, $\tilde{F}_n(.)$ denotes the n-fold Stieltjes convolution of $\tilde{F}(.)$, i.e. $\tilde{F}_1(x) = \tilde{F}(x)$, and $\tilde{F}_n(x) = \int_0^x \tilde{f}(y) \cdot \tilde{F}_{n-1}(x-y)\, dy$ for all $n > 1$:

$$\xi_u^0(m,n) = \binom{m+n}{m} \cdot \overline{F}(\theta)^m \cdot F(\theta)^n \cdot \left\{ \begin{array}{c} \tilde{F}_n(u - m \cdot \theta) - \\ F(\theta) \cdot \tilde{F}_{n+1}(u - m \cdot \theta) - \\ \overline{F}(\theta) \cdot \tilde{F}_n(u - (m+1) \cdot \theta) \end{array} \right\}$$

The first term between curly brackets reflects the probability that the first m preventive and n corrective maintenance actions are completed within the first u units of cumulative up time. Similarly, the second and third term reflect the probability that the next i.e. $m+n+1^{st}$ maintenance action, which is preventive with probability $\overline{F}(\theta)$ and corrective with probability $F(\theta)$, is also completed within the remaining up time. Together, these terms denote the probability that the first u units of cumulative up time are attained somewhere between the $m+n^{st}$ and the $m+n+1^{st}$ maintenance action, provided that the first m maintenance actions are preventive and the following n maintenance actions are corrective. For notational convenience, and without loss of generality, we will use the notation of $\xi_u^0(m,n)$ in deriving analytical expressions for $P_0(T_u \leq t)$ in the sequel.

4.3.3 Stochastic repair and stochastic maintenance times

Given the number of preventive maintenance actions m and corrective maintenance actions n, observed with probability $\xi_u^0(m,n)$, the corresponding down times do not accumulate to more than $t-u$ time units with probability $G_m \circ H_n(t-u)$. Here,

$G \circ H(x) = \int_0^x g(y) \cdot H(x-y)\, dy$ denotes the well-known Stieltjes convolution for computing the sum of independent stochastic variables. Summarizing, this yields the following expression for $P_0(T_u \leq t)$:

$$P_0(T_u \leq t) = \sum_{m\cdot\theta < u \leq (m+n+1)\cdot\theta} \xi_u^0(m,n) \cdot G_m \circ H_n(t-u)$$

Following a similar argument, the first and higher moments of T_u can be derived in a rather straightforward manner, as long as the corresponding moments of the maintenance and repair time distributions are available. For example, the first two moments of $T_u - u$, given that the system starts with an up time at time $t = 0$, are determined as follows:

$$E_0\{T_u - u\} = \sum_{m\cdot\theta < u \leq (m+n+1)\cdot\theta} \xi_u^0(m,n) \cdot \{m \cdot \mu_P + n \cdot \mu_R\}$$

$$E_0\{(T_u - u)^2\} = \sum_{m\cdot\theta < u \leq (m+n+1)\cdot\theta} \xi_u^0(m,n) \cdot \left\{m \cdot \sigma_P^2 + n \cdot \sigma_R^2 + (m \cdot \mu_P + n \cdot \mu_R)^2\right\}$$

From the transparancy of these expressions, it is immediately clear that the calculation of $P_0(T_u \leq t)$ for different values of t, is to be preferred above the calculation of $P_0(U_t \leq u)$ for different values of u, at least in view of the inherent mathematical complications. Nevertheless, mathematical models for the interval availability distribution have always been formulated in terms of U_t, rather than T_u, at least up to our knowledge of existing literature. The main reason for this is that the interval availability distribution is formally defined as $P_0(U_t \leq u)$, and not as $P_0(T_u \leq t)$. To avoid confusion, we will refer to $P_0(T_u \leq t)$ as the availability interval distribution, rather than the interval availability distribution.

4.3.4 Stochastic repair and deterministic maintenance times

If the time required for preventive maintenance is fixed ($\sigma_P = 0$), the number of preventive maintenance actions m must satisfy $m \cdot \mu_P \leq t - u$. In that case, $G_m \circ H_n(t-u) = H_n(t-u-m\cdot\mu_P)$, and thus an alternative expression can be derived for $P_0(T_u \leq t)$:

$$P_0(T_u \leq t) = \sum_{\substack{m\cdot\theta < u \leq (m+n+1)\cdot\theta \\ m\cdot\mu_P \leq t-u}} \xi_u^0(m,n) \cdot H_n(t-u-m\cdot\mu_P)$$

If no failures occur during the first u units of cumulative up time, the number of preventive maintenance actions equals $\lceil u/\theta - 1 \rceil$ with probability one. Moreover,

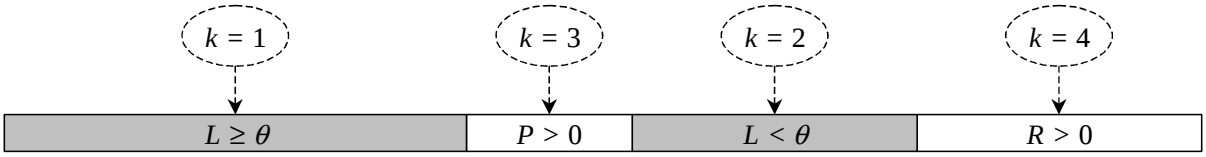


Figure 4.3: Possible states of the production system: preventive up time ($k = 1$), corrective up time ($k = 2$), preventive down time ($k = 3$) and corrective down time ($k = 4$).

we know for sure that the corresponding down times accumulate to $\lceil u/\theta - 1 \rceil \cdot \mu_P$ time units. Since each maintenance action is preventive with probability $\overline{F}(\theta)$ and corrective with probability $F(\theta)$, this yields the following additional and a priori information with respect to the cumulative distribution function $P_0(T_u \leq t)$:

$$P_0(T_u = u + \lceil u/\theta - 1 \rceil \cdot \mu_P) = \overline{F}(\theta)^{\lceil u/\theta - 1 \rceil} \cdot \overline{F}(u - \lceil u/\theta - 1 \rceil \cdot \theta)$$

A typical example of deterministic maintenance times, but stochastic repair times, can be found in the replacement of single components and/or complete (sub)systems. In general, preventive replacements require a fixed amount of time, since they are perfectly plannable. Corrective replacements, however, often require an additional waiting time, since the required resources are not always readily available on request.

4.4 Limiting behavior of the system

In this section, we consider the case where the production system starts in an arbitrary state at time $t = 0$, and preventive maintenance is carried out as soon as the system has been operational (up) for $\theta > 0$ time units. As a starting point of our analysis, we observe that the following situations can occur (see Figure 4.3):

(1) the system starts in a preventive up time,

(2) the system starts in a corrective up time,

(3) the system starts in a preventive down time,

(4) the system starts in a corrective down time.

Here, a preventive (corrective) up time is defined as a continuous period of time during which the system is operational (up), and which is **terminated** by a preventive

(corrective) maintenance action. This distinction will appear to be convenient when deriving expressions for the availabity interval distribution. Similarly, a preventive (corrective) down time is defined as a continuous period of time during which the system is not operational (down), and which is **initiated** by a preventive (corrective) maintenance action. If we denote with $\tilde{\mu}_L = \int_0^\theta 1 - \tilde{F}(\tau)\, d\tau$ the mean length of a corrective up time, it is easily verified that the corresponding limiting probabilities π_1, π_2, π_3 and π_4 are interrelated as follows. Here, π_k denotes the long run average fraction of time that the system remains in state k:

$$\pi_1 : \pi_2 : \pi_3 : \pi_4 = \theta \cdot \overline{F}(\theta) : \tilde{\mu}_L \cdot F(\theta) : \mu_P \cdot \overline{F}(\theta) : \mu_R \cdot F(\theta)$$

Together with the normalization condition $\sum_k \pi_k = 1$, this yields the required and unique values for π_k $(1 \leq k \leq 4)$. Let us now denote with $P_k(T_u \leq t)$ the probability of at least u units of cumulative up time during an arbitrary interval of t time units, given that the system starts in state k. Then obviously, the probability $P_\infty(T_u \leq t)$ of at least u units of cumulative up time during an arbitrary interval of t time units, given that the system starts in a stationary state, is given by:

$$P_\infty(T_u \leq t) = \sum_{k=1}^{4} \pi_k \cdot P_k(T_u \leq t)$$

In a similar way, the first two moments $E_\infty\{T_u - u\}$ and $E_\infty\{(T_u - u)^2\}$ of $T_u - u$ can be determined. In the following sections, we will elaborate on these expressions in more detail.

4.4.1 Start in preventive up time ($k = 1$)

If the system starts in state $k = 1$, the remaining preventive up time is described by a random variable $R_1 \in [0, \theta]$ with cumulative distribution function $\Phi_1(.)$:

$$\Phi_1(\tau) = P(R_1 \leq \tau) = \frac{\tau}{\theta} \quad , \quad 0 \leq \tau \leq \theta$$

As a starting point, let us determine the probabilities $\xi_u^1(m, n)$ of exactly m preventive and n corrective maintenance actions during the first u units of cumulative up time, given that the system start in a preventive up time $(k = 1)$. First of all, we observe that no maintenance occurs if the remaining up time exceeds the amount of u time units. In all other cases, the first maintenance action must be preventive by assumption. In formula, this yields $\xi_u^1(0, 0) = 1 - \Phi_1(u)$, and $\xi_u^1(0, n) = 0$ for all $n \geq 1$. For all other values of $m \geq 1$ and $n \geq 0$, $\xi_u^1(m, n)$ yields an expression which is similar to $\xi_u^0(m, n)$:

$$\xi_u^1(m,n \mid m \geq 1) = \tbinom{m+n-1}{m-1} \cdot \overline{F}(\theta)^{m-1} \cdot F(\theta)^n \cdot \left\{ \begin{array}{c} \Phi_1 \circ \tilde{F}_n(u-(m-1)\cdot\theta) - \\ F(\theta) \cdot \Phi_1 \circ \tilde{F}_{n+1}(u-(m-1)\cdot\theta) - \\ \overline{F}(\theta) \cdot \Phi_1 \circ \tilde{F}_n(u-m\cdot\theta) \end{array} \right\}$$

Given the number of preventive maintenance actions m and corrective maintenance actions n, the corresponding down times do not accumulate to more than $t-u$ time units with probability $G_m \circ H_n(t-u)$. As a consequence, we arrive at the following expression for $P_1(T_u \leq t)$:

$$P_1(T_u \leq t) = \sum_{m,n} \xi_u^1(m,n) \cdot G_m \circ H_n(t-u)$$

4.4.2 Start in corrective up time ($k = 2$)

If the system starts in state $k = 2$, the remaining corrective up time is described by a random variable $R_2 \in [0, \theta]$ with cumulative distribution function $\Phi_2(.)$:

$$\Phi_2(\tau) = P(R_2 \leq \tau) = \frac{1}{\tilde{\mu}_L} \cdot \int_0^\tau 1 - \tilde{F}(\upsilon)\, d\upsilon \quad , \quad 0 \leq \tau \leq \theta$$

In a similar way, we can determine the probabilities $\xi_u^2(m,n)$ of exactly m preventive and n corrective maintenance actions during the first u units of cumulative up time, given that the system start in a preventive up time ($k = 2$). Since the first maintenance action must be corrective by assumption, we find $\xi_u^2(0,0) = 1 - \Phi_2(u)$, and $\xi_u^2(m,0) = 0$ for all $m \geq 1$. For all other values of $m \geq 0$ and $n \geq 1$, $\xi_u^2(m,n)$ yields an expression which is similar to $\xi_u^0(m,n)$ and $\xi_u^1(m,n)$:

$$\xi_u^2(m,n \mid n \geq 1) = \tbinom{m+n-1}{n-1} \cdot \overline{F}(\theta)^m \cdot F(\theta)^{n-1} \cdot \left\{ \begin{array}{c} \Phi_2 \circ \tilde{F}_{n-1}(u-m\cdot\theta) - \\ F(\theta) \cdot \Phi_2 \circ \tilde{F}_n(u-m\cdot\theta) - \\ \overline{F}(\theta) \cdot \Phi_2 \circ \tilde{F}_{n-1}(u-(m+1)\cdot\theta) \end{array} \right\}$$

Given the number of preventive maintenance actions m and corrective maintenance actions n, the corresponding down times do not accumulate to more than $t-u$ time units with probability $G_m \circ H_n(t-u)$. Obviously, this yields the following expression for $P_2(T_u \leq t)$:

$$P_2(T_u \leq t) = \sum_{m,n} \xi_u^2(m,n) \cdot G_m \circ H_n(t-u)$$

Of course, our analysis leads to similar expressions for $P_1(T_u \leq t)$ and $P_2(T_u \leq t)$, since their only difference orginates from the remaining up times, with corresponding

distribution functions $\Phi_1(.)$ and $\Phi_2(.)$. Note that combination of $P_1(T_u \leq t)$ and $P_2(T_u \leq t)$ would not simplify these expressions. In fact, the distinction between preventive and corrective up times certainly facilitated their derivation.

4.4.3 Start in preventive down time ($k = 3$)

If the system starts in state $k = 3$, the remaining preventive down time is described by a random variable $R_3 \in [0, \infty\rangle$ with cumulative distribution function $\Phi_3(.)$:

$$\Phi_3(\tau) = P(R_3 \leq \tau) = \frac{1}{\mu_P} \cdot \int_0^\tau 1 - G(v)\ dv \quad , \quad \tau \geq 0$$

As a starting point, we observe that the first up time starts as soon as preventive maintenance is completed. Therefore, $\xi_u^0(m, n)$ denotes the probability of exactly m preventive and n corrective maintenance actions during the first u units of cumulative up time. Given the number of preventive maintenance actions m and corrective maintenance actions n, the corresponding down times do not accumulate to more than $t - u$ time units with probability $\Phi_3 \circ G_m \circ H_n(t - u)$. This yields the following expression for $P_3(T_u \leq t)$:

$$P_3(T_u \leq t) = \sum_{m,n} \xi_u^0(m, n) \cdot \Phi_3 \circ G_m \circ H_n(t - u)$$

4.4.4 Start in corrective down time ($k = 4$)

If the system starts in state $k = 4$, the remaining corrective down time is described by a random variable $R_4 \in [0, \infty\rangle$ with cumulative distribution function $\Phi_4(.)$:

$$\Phi_4(\tau) = P(R_4 \leq \tau) = \frac{1}{\mu_R} \cdot \int_0^\tau 1 - H(v)\ dv \quad , \quad \tau \geq 0$$

In a similar way, we observe that the first up time starts as soon as corrective maintenance is completed. Given the number of preventive maintenance actions m and corrective maintenance actions n, with probability $\xi_u^0(m, n)$, the corresponding down times do not accumulate to more than $t - u$ time units with probability $\Phi_4 \circ G_m \circ H_n(t - u)$. This yields the following expression for $P_4(T_u \leq t)$:

$$P_4(T_u \leq t) = \sum_{m,n} \xi_u^0(m, n) \cdot \Phi_4 \circ G_m \circ H_n(t - u)$$

Once again, these expressions for $P_3(T_u \leq t)$ and $P_4(T_u \leq t)$ are very similar, since their only difference originates from the remaining down times, with corresponding distribution functions $\Phi_3(.)$ and $\Phi_4(.)$.

4.5 The optimal maintenance interval

So far, we have considered the preventive maintenance interval θ to be a given constant. In this section, we will present a rather straightforward algorithm with which an optimal preventive maintenance interval θ^* can be determined from an interval availability point of view. To this end, a plausible choice for an objective function is presented first. Subsequently, the objective function under consideration is evaluated for a production system with Gamma distributed repair and fixed maintenance times. In that case, explicit formulas can be derived, which strongly reduce the complexity of the optimization problem.

4.5.1 Objective functions

In classical maintenance theory, an optimal preventive maintenance interval $\theta_0 < \infty$ for a production system is usually determined by maximizing its limiting availability A_∞ (see section 4.2). In our setting here, a similar approach would be to minimize the expected time $E_\infty\{T_u\}$ required to attain a cumulative up time of u time units, given that the system start in an arbitrary state at time $t = 0$. In general, these objectives are not equivalent, i.e. $E_\infty\{T_u\} \cdot A_\infty \neq u$, in particular if u is relatively small compared to the expected lifetime of the production system. Anyhow, none of these objectives accounts for the fact that $Var\{T_u\}$ is also a value of great interest, since it provides information about the short term behavior of the production system. As an alternative, we have chosen to minimize the **ω-percentile** of the availability interval T_u, where $0 < \omega < 1$ is a user-defined constant. This is quite natural, since it provides information about the one-sided confidence interval for the required time to complete a cumulative workload of u time units. In line with this, our objective becomes to minimize $f_u^\omega(\theta)$, with parameters u and ω:

$$f_u^\omega(\theta) = \inf\{t \geq u \mid P_\infty(T_u \leq t \mid \theta) \geq \omega\}$$

4.5.2 Function evaluation

Unfortunately, the evaluation of $f_u^\omega(\theta)$ for a given value of θ is rather complicated from a mathematical point of view. As a starting point, numerical approximations for the convolutions $\tilde{F}_n$, $\Phi_1 \circ \tilde{F}_n$, and $\Phi_2 \circ \tilde{F}_n$ have to be calculated, in order to determine $\xi_u^0(m,n)$, $\xi_u^1(m,n)$ and $\xi_u^2(m,n)$ for all $m, n \geq 0$. In this respect, an upper bound M resp. N on the number of preventive resp. corrective maintenance actions m resp. n needs to be identified, in order to truncate the infinite summations appearing in

$P_k(T_u \leq t)$. For $k = 0$, this can be done in a rather straightforward manner, if one realizes that the following relation holds for all $M, N \geq 0$:

$$P_0(T_u \leq t) - \sum_{m=0}^{M} \sum_{n=0}^{N} \xi_u^0(m,n) \cdot G_m \circ H_n(t-u) \leq 1 - \sum_{m=0}^{M} \sum_{n=0}^{N} \xi_u^0(m,n)$$

In other words, M and N can be increased in a stepwise manner, until this restriction is satisfied. Obviously, similar results can be obtained for other values of k, which yields the desired result. Subsequently, the stationary probabilities π_k $(1 \leq k \leq 4)$, as well as the convolutions $\Phi_3 \circ G_m \circ H_n$ and $\Phi_4 \circ G_m \circ H_n$, have to be numerically approximated in order to evaluate $P_\infty(T_u \leq t)$ for a given value of t. Finally, a one-dimensional search procedure (e.g. bi-section) has to be carried out in order to identify the smallest value of t for which $P_\infty(T_u \leq t) \geq \omega$. In general, i.e. for arbitrary distribution functions $F(.)$, $G(.)$ and $H(.)$, this yields a complex procedure which requires a large amount of computational effort. Under some special conditions, however, the complexity of evaluating $f_u^\omega(\theta)$ can be reduced significantly, in particular if repair times are Gamma distributed random variables, and preventive maintenance times are fixed.

4.5.3 Gamma distributed repair and fixed maintenance times

In this section, we will restrict ourselves to the case where repair times are Gamma distributed random variables with parameters $\alpha = \mu_R^2 \cdot \sigma_R^{-2}$ and $\beta = \mu_R^{-1} \cdot \sigma_R^2$, and preventive maintenance requires a fixed amount of time $\mu_P > 0$ (i.e. $\sigma_P = 0$). Under these assumptions, explicit formulas can be derived for $G_m \circ H_n$, $\Phi_3 \circ G_m \circ H_n$ and $\Phi_4 \circ G_m \circ H_n$, which appear in the definitions of $P_k(T_u \leq t)$. As a starting point, we define $\Psi_{\alpha,\beta}(.)$ for notational convenience:

$$\Psi_{\alpha,\beta}(\tau) \equiv \int_0^\tau \Gamma_{\alpha,\beta}(\upsilon)\, d\upsilon = \tau \cdot \Gamma_{\alpha,\beta}(\tau) - \alpha \cdot \beta \cdot \Gamma_{\alpha+1,\beta}(\tau)$$

Lemma 5 *If repair times are Gamma distributed random variables with parameters α and β, i.e. $\mu_R = \alpha \cdot \beta$ and $\sigma_R^2 = \alpha \cdot \beta^2$, and preventive maintenance requires a fixed amount of time $\mu_P > 0$, then $G_m \circ H_n(t-u)$, $\Phi_3 \circ G_m \circ H_n(t-u)$ and $\Phi_4 \circ G_m \circ H_n(t-u)$ can be derived analytically by means of the following explicit formulas:*

$$G_m \circ H_n(t-u) = \Gamma_{n\cdot\alpha,\beta}(t - u - m \cdot \mu_P)$$

$$\Phi_3 \circ G_m \circ H_n(t-u) = \frac{\Psi_{n\cdot\alpha,\beta}(t - u - m \cdot \mu_P) - \Psi_{n\cdot\alpha,\beta}(t - u - (m+1) \cdot \mu_P)}{\mu_P}$$

$$\Phi_4 \circ G_m \circ H_n(t-u) = \frac{\Psi_{n\cdot\alpha,\beta}(t-u-m\cdot\mu_P) - \Psi_{(n+1)\cdot\alpha,\beta}(t-u-m\cdot\mu_P)}{\mu_R}$$

For the proof of this theorem, we refer to the appendix. Here, we only mention that efficient computer programming codes are available for the calculation of Gamma distributions, e.g. see Temme (1994). Hence, the only convolutions that need to be numerically approximated are $\tilde{F}_n$, $\Phi_1 \circ \tilde{F}_n$, and $\Phi_2 \circ \tilde{F}_n$, which appear in the definitions of $\xi_u^0(m,n)$, $\xi_u^1(m,n)$, and $\xi_u^2(m,n)$. Since $\tilde{F}(\theta) = \Phi_1(\theta) = \Phi_2(\theta) = 1$ by definition, and thus $\tilde{F}(.)$, $\Phi_1(.)$ and $\Phi_2(.)$ all have finite support, these convolutions can be determined to a sufficient level of detail within reasonable computational times.

In this respect, another but intuitively less attractive possibility, is to model both preventive and corrective maintenance times as Gamma distributed random variables with the same shape parameter β. This would imply, however, that either average preventive maintenance times are larger than average corrective maintenance times ($\mu_P > \mu_R$), or the coefficient of variation of preventive maintenance times is larger than the coefficient of variation of corrective maintenance times ($\sigma_P/\mu_P > \sigma_R/\mu_R$). Since none of these alternatives is likely to occur in practice, these assumptions would leave us with a theoretical exercise of almost no practical relevance.

4.5.4 Optimization algorithm

Our optimization algorithm starts with observing that $f_u^{\omega}(\theta)$ has an infinite number of discontinuities of the form $\theta = u/k$ ($k \geq 1$), because the number of preventive maintenance actions equals $\lfloor u/\theta \rfloor$ or $\lceil u/\theta \rceil$ if no failures occur during the required interval of u units cumulative up time. Therefore, and to avoid the risk of sub-optimization, we decomposed our global optimization procedure into a series of consecutive local optimization procedures within disjunct ranges of the form $[u/(k+1), u/k\rangle$. As a starting point, however, we determine the optimal maintenance interval $\theta_0 < \infty$ from a limiting availability point of view (Barlow and Proschan 1965), and assume that $\theta^* \leq \theta_0$. In line with this, our first range under consideration becomes $[u/k_0, \theta_0\rangle$, where $k_0 = \lceil u/\theta_0 \rceil$.

Our optimization algorithm is now based on the assumption that $f_u^{\omega}(\theta)$ is a piecewise unimodal function within each of the above-mentioned ranges. As a starting point, we determine the optimal preventive maintenance interval $\theta_{k_0}^*$ within the range $[u/k_0, \theta_0\rangle$ with the use of golden section search (Brent 1973). In a similar way, and starting with $k = k_0$, we determine the optimal maintenance interval θ_k^* within the range $[u/(k+1), u/k\rangle$, and at the same time keep track of the best-so-far mainte-

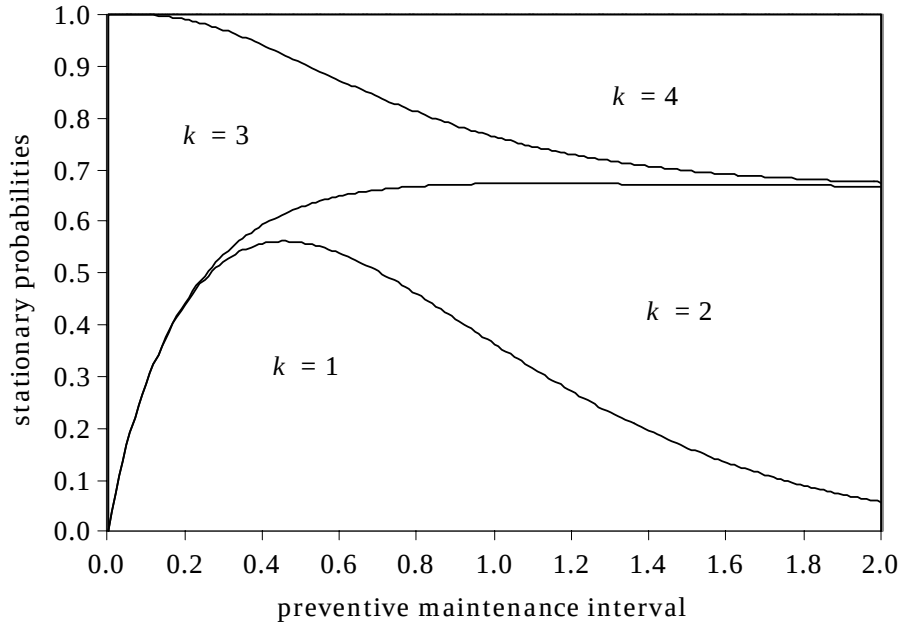


Figure 4.4: Stationary probabilities π_k in relation to the preventive maintenance interval θ, for a production system with Gamma distributed life times ($\mu_L = 1$, $\sigma_L = \frac{1}{2}$), Gamma distributed repair times ($\mu_R = \frac{1}{2}$, $\sigma_R = \frac{1}{4}$) and fixed preventive maintenance times ($\mu_P = \frac{1}{4}$, $\sigma_P = 0$).

nance interval θ^* within the range $[u/(k+1), \theta_0\rangle$. As soon as the optimal maintenance interval θ^* does not change in two consecutive iterations, the algorithm is terminated. Obviously, this stop criterion is based on the underlying assumption that $f_u^\omega(\theta_{k+1}^*) \geq f_u^\omega(\theta_k^*)$ implies that $\theta^* \geq \theta_{k+1}^*$.

Under some weak conditions, it can be shown that this procedure would lead to the optimal preventive maintenance interval, if we were concerned with the limiting availability of the production system (Barlow and Proschan 1965). Unfortunately, we have not been able to prove this results for the availability interval distribution. On the other hand, we have had no indications so far that these assumptions strongly affect the performance of our numerical optimization algorithm. Anyhow, the computational results that will be presented in the following sections should be interpreted as a lower bound for the savings that can be obtained if the optimal maintenance interval is determined from an interval rather than a limiting availability perspective.

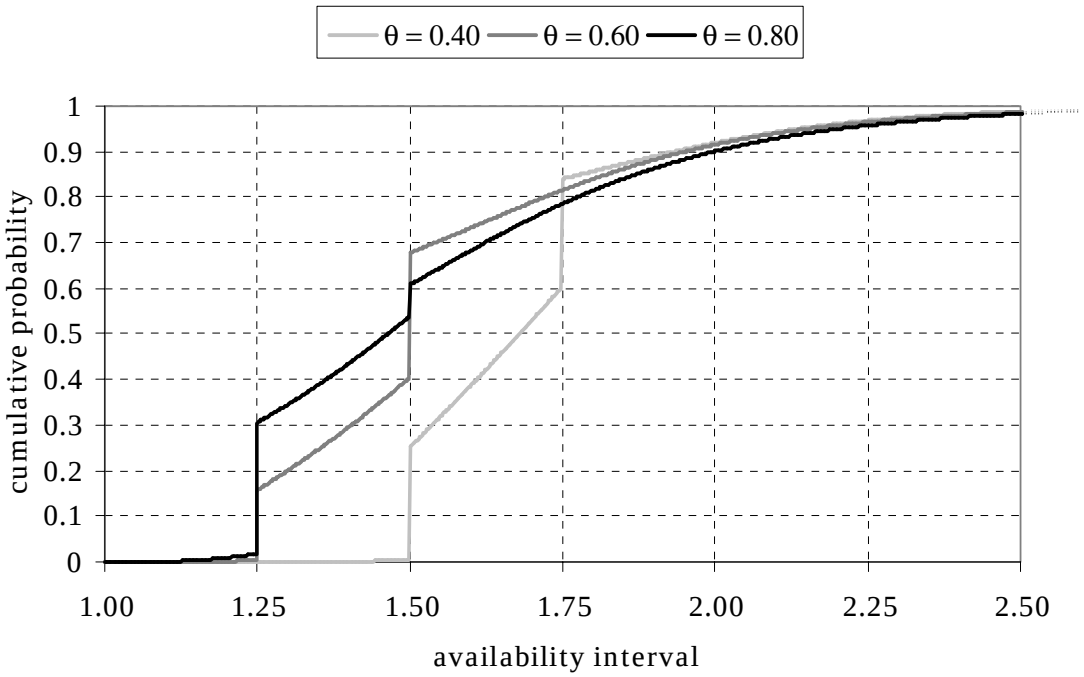


Figure 4.5: Availability interval distribution $P_\infty(T_u \leq t)$, in case $u = 1$, for a production system with Gamma distributed life times ($\mu_L = 1$, $\sigma_L = \frac{1}{2}$), Gamma distributed repair times ($\mu_R = \frac{1}{2}$, $\sigma_R = \frac{1}{4}$), fixed preventive maintenance times ($\mu_P = \frac{1}{4}$, $\sigma_P = 0$), and different preventive maintenance intervals θ.

4.6 Numerical example

Let us now present a numerical example in order to illustrate the above-mentioned methods in more detail. To this end, we consider a production system with Gamma distributed life times ($\mu_L = 1$ and $\sigma_L = \frac{1}{2}$), Gamma distributed repair times ($\mu_R = \frac{1}{2}$ and $\sigma_R = \frac{1}{4}$) and fixed preventive maintenance times ($\mu_P = \frac{1}{4}$ and $\sigma_P = 0$). Moreover, we assume that the required cumulative up time equals $u = 1$ time units, which is exactly equal to the expected life time of the production system.

4.6.1 Stationary probabilities

As a starting point, we determine the stationary probabilities π_k of starting in state k, where $1 \leq k \leq 4$ (see Figure 4.4). As can be seen from this figure, π_3 tends to one as θ tends to zero. In that case, the production system is maintained preventively all the time, and the system is always down for preventive maintenance. For similar reasons, both π_1 and π_3 tend to zero if θ tends to infinity. In that case, all maintenance will be

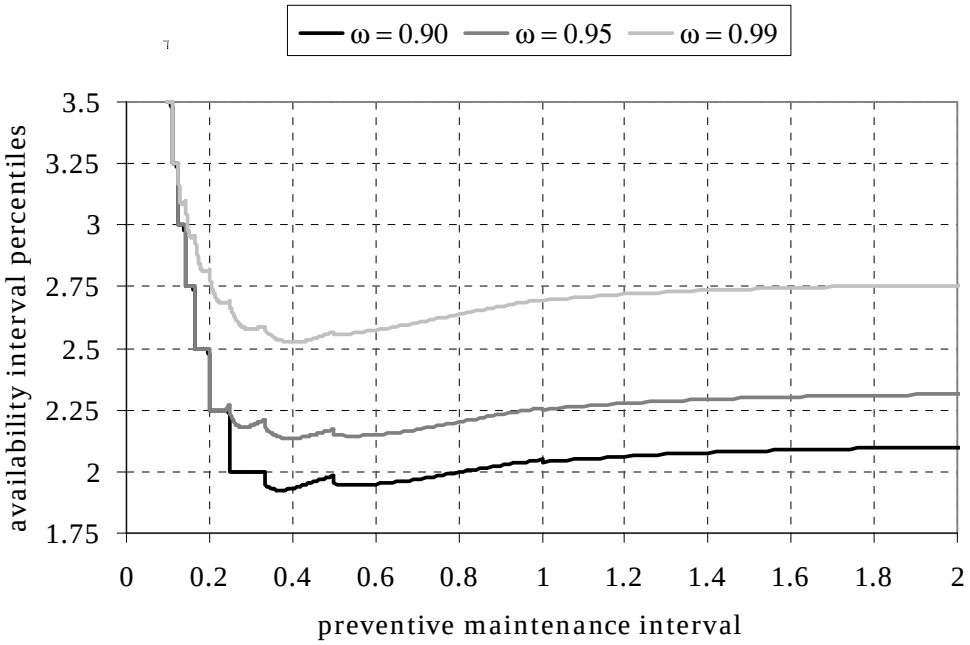


Figure 4.6: Percentiles of the availability interval T_u, in case $u = 1$, in relation to the preventive maintenance interval θ, for a production system with Gamma distributed life times ($\mu_L = 1$, $\sigma_L = \frac{1}{2}$), Gamma distributed repair times ($\mu_R = \frac{1}{2}$, $\sigma_R = \frac{1}{4}$) and fixed preventive maintenance times ($\mu_P = \frac{1}{4}$, $\sigma_P = 0$).

corrective, and the system will be either up or down, with corresponding probabilities $\pi_2 : \pi_4 = \mu_L : \mu_R = 2 : 1$. Obviously, this corresponds to a limiting availability of $A_\infty = \mu_L/(\mu_L + \mu_R) = \frac{2}{3}$. From Figure 4.4, it can also be concluded that the optimal maintenance interval $\theta_0 \approx 1.10$ if our objective is to maximize the limiting availability $A_\infty = \pi_1 + \pi_2$ of the production system. Moreover, this optimal maintenance interval hardly outperforms a corrective maintenance strategy ($\theta^* \to \infty$).

4.6.2 Availability interval distribution

To continue our analysis, let us determine the limiting behavior of the production system for different values of θ, in terms of the cumulative distribution function $P_\infty(T_u \leq t)$. In this particular example, we determined these probabilities for $\theta = \frac{4}{5}$, $\theta = \frac{3}{5}$, and $\theta = \frac{2}{5}$ respectively (see Figure 4.5). Along the horizontal axis, the discontinuities $u + \lfloor u/\theta \rfloor \cdot \mu_P$ and $u + \lceil u/\theta \rceil \cdot \mu_P$ for T_u are clearly visible. Moreover, there are some strong indications that the optimal maintenance interval is closely related to the desired confidence interval. To be specific, the best maintenance interval

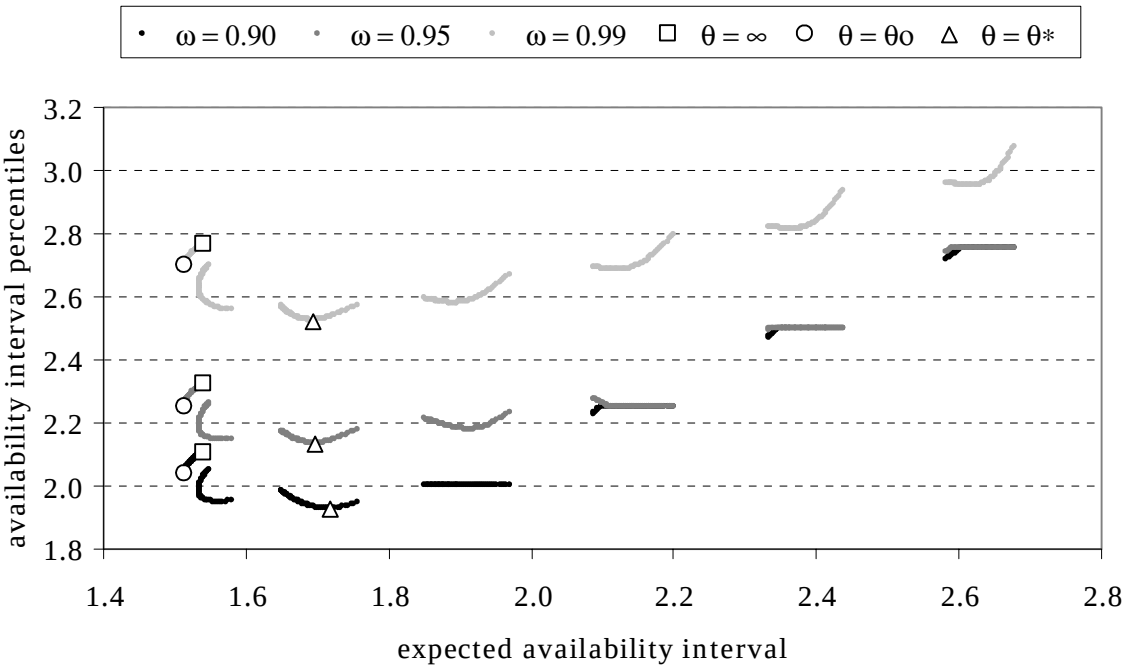


Figure 4.7: Expectation $E_\infty\{T_u\}$ versus ω-percentiles $f_u^\omega(\theta)$ of the availability interval T_u, in case $u = 1$, for a production system with Gamma distributed life times ($\mu_L = 1$, $\sigma_L = \frac{1}{2}$), Gamma distributed repair times ($\mu_R = \frac{1}{2}$, $\sigma_R = \frac{1}{4}$), fixed preventive maintenance times ($\mu_P = \frac{1}{4}$, $\sigma_P = 0$), and different preventive maintenance intervals θ.

equals $\theta = \frac{4}{5}$ for $\omega = \frac{1}{2}$, $\theta = \frac{3}{5}$ for $\omega = \frac{3}{4}$, and $\theta = \frac{2}{5}$ for $\omega = \frac{7}{8}$. Apparently, the optimal maintenance interval decreases if the guaranteed performance during a finite period of time (interval availability) becomes more important than the average performance during an infinite period of time (limiting availability).

4.6.3 Optimal maintenance interval

In order to arrive at the optimal maintenance interval θ^*, we determine $f_u^\omega(\theta)$ for different values of both $\omega \in \{0.90, 0.95, 0.99\}$ and $\theta \in \langle 0, 2]$, where $u = 1$. The results are depicted in Figure 4.6. A closer look at these results provided the following optimal maintenance intervals: $\theta^* = 0.38$ for $\omega = 0.90$, $\theta^* = 0.40$ for $\omega = 0.95$ and $\theta^* = 0.40$ for $\omega = 0.99$. Apparently, the optimal maintenance interval θ^* hardly depends on the value of ω in this particular example. Nevertheless, we can conclude from Figure 4.6 that the optimal maintenance interval for interval availability $\theta^* \approx 0.40$ is significantly smaller than the optimal maintenance interval $\theta_0 \approx 1.10$ from a

limiting availability point of view. In the following section, we will carry out a series of numerical experiments in order to investigate the relation between θ^* and θ_0 on the one hand, and between $f_u^\omega(\theta^*)$ and $f_u^\omega(\theta_0)$ on the other hand.

4.6.4 Limiting versus interval availability

Let us now further elaborate upon the difference between the average and guaranteed performance of the production system. To this end, we compared the expectation $E_\infty\{T_1\}$ and the ω-percentiles $f_1^\omega(\theta) = \inf\{t \geq 1 \mid P_\infty\{T_1 \leq t\} \geq \omega\}$ of the time T_1 required to attain a cumulative up time of exactly $u = 1$ time unit. Of course, this was done for different values of θ and $\omega \in \{0.90, 0.95, 0.99\}$. The results are depicted in Figure 4.7. As we expected, the discontinuities of the form $\theta = u/k$ are clearly visible, and cause empty spaces in this figure. Moreover, we observe that a corrective maintenance strategy ($\theta = \infty$) performs poor in both dimensions. Starting from here, decreasing the maintenance interval leads to an improvement in both dimensions, up to the point where the expected value $E_\infty\{T_u\}$ is minimized ($\theta \approx \theta_0$). Subsequently, reducing the maintenance interval leads to degradations in the first dimension, but at the same time to further improvements in the second dimension, up to the point where the ω-percentile $f_u^\omega(\theta)$ is minimized ($\theta = \theta^*$). At this point, further reductions in the preventive maintenance interval leads to degradations in both dimensions.

4.7 Computational results

In this section, we will present the results of a series of numerical experiments that were carried out for a production system with Gamma distributed lifetimes, Gamma distributed repair times, and fixed preventive maintenance times. Amongst other factors, the main objectives of these numerical experiments were (i) to determine what happens if the optimal maintenance interval is determined from an interval availability rather than a limiting availability point of view, and (ii) to investigate how the optimal maintenance interval for interval availability depends on the characteristics of the production system.

For notational convenience, and without loss of generality, we assumed that $\mu_L = 1$ in each test problem. Moreover, the relevant parameters μ_R/μ_L, μ_P/μ_R, σ_L/μ_L and σ_R/μ_R were varied between $\frac{1}{2}$ and $\frac{1}{4}$, in order to arrive at 16 production systems with - at least theoretically - different short term behavior. For each production system, we generated a total of 9 test problems by choosing $u \in \{1, 2, 3\}$ and $\omega \in \{0.90, 0.95, 0.99\}$. For each of the 144 test problems obtained this way, the optimal

Table 4.1: Comparison of the optimal maintenance intervals θ^* for interval availability and θ_0 for limiting availability, as well as the corresponding ω-percentiles for T_u, for different values of u and ω.

		$\{1-\theta_0/\theta^*\}\times 100\%$			$\{1-f_u^\omega(\theta^*)/f_u^\omega(\theta_0)\}\times 100\%$		
		minimal	average	maximal	minimal	average	maximal
	$u=1$	20.6	41.2	67.6	0.0	6.7	17.8
$\omega=0.90$	$u=2$	18.3	29.6	45.4	0.7	3.4	9.4
	$u=3$	19.5	26.1	35.9	0.4	2.4	6.7
	$u=1$	20.6	39.7	64.6	1.9	10.9	20.8
$\omega=0.95$	$u=2$	10.6	32.6	46.4	0.8	5.5	14.7
	$u=3$	19.5	29.9	37.6	0.6	3.4	9.7
	$u=1$	35.6	49.1	63.7	2.3	12.1	29.0
$\omega=0.99$	$u=2$	32.8	43.4	49.9	1.3	7.4	21.0
	$u=3$	18.4	36.7	46.5	0.7	4.7	13.6

maintenance interval θ^* for interval availability, the optimal maintenance interval θ_0 for limiting availability, and the corresponding availability interval percentiles $f_u^\omega(\theta^*)$ and $f_u^\omega(\theta_0)$ were determined with the use of our optimizatin algorithm. An overview of all test problems is depicted in Table 4.1. In addition, the results for all test problems with $u=1$ are depicted in Table 4.2.

As a starting point, it is easily verified from Table 4.1 that significant improvements can be obtained in the short term behavior of a production system, if the optimal maintenance interval is determined from an interval availability rather than a limiting availability point of view. Depending on the required amount of cumulative up time $u>0$, the required percentile $\omega<1$, and the characteristics of the production system, the corresponding improvements are substantial, with a maximum of about 30% in the availability interval. To achieve this, a 10% to 70% reduction in the preventive maintenance interval was typical.

Summarizing, the general conclusion that can be drawn from Table 4.1, is that the required up time $u>0$ and percentile $\omega<1$ on the one hand, and the optimal maintenance intervals $\theta^*$and θ_0 with availability interval percentiles $f_u^\omega(\theta^*)$ and $f_u^\omega(\theta_0)$ on the other hand, are interrelated as follows:

- an increase in the desired up time u usually goes together with an increase in the optimal maintenance interval θ^* for interval availability, as well as a decrease in the relative performance of θ^* versus θ_0 in terms of $f_u^\omega(\theta_0)/f_u^\omega(\theta^*)$;

- an increase in the desired percentile ω usually goes together with a decrease in the optimal maintenance interval θ^* for interval availability, as well as an increase in the relative performance of θ^* versus θ_0 in terms of $f_u^\omega(\theta_0)/f_u^\omega(\theta^*)$.

In a similar way, it can be derived from Table 4.2 to which extent the characteristics of the production system affect the optimal maintenance intervals θ^*, as well as the corresponding availability interval percentiles $f_u^\omega(\theta^*)$. As a starting point, and as expected, we observe that θ_0 depends on σ_L/μ_L and μ_P/μ_R only. In addition, the following observations were made from Table 4.2:

- an increase in the ratio of repair versus maintenance times usually goes together with a decrease in the optimal maintenance interval θ^* for interval availability, as well as a decrease in the corresponding availability interval percentile $f_u^\omega(\theta^*)$;

- an increase in the variation of life and/or repair times usually goes together with a decrease in the optimal maintenance interval θ^* for interval availability, as well as an increase in the corresponding availability interval percentile $f_u^\omega(\theta^*)$.

Although intuitively attractive, these rules of thumbs do not cover all possible situations that may occur, e.g. see $\mu_R/\mu_L = \mu_P/\mu_R = \frac{1}{2}$ and $\sigma_L/\mu_L = \sigma_R/\mu_R = \frac{1}{4}$ in Table 4.2. Nevertheless, we conclude that the guaranteed availability interval of a production system can be improved significantly, if the optimal preventive maintenance interval is determined from an interval availability perspective. From a practical point of view, this means that preventive maintenance is a powerful instrument to increase the **controllability** or **predictability** of a production system.

4.8 Concluding remarks

In this paper, we have presented a series of mathematical models which can be used to determine the availability interval distribution for a production system which is maintained preventively at regular intervals, according to an age replacement strategy. Moreover, we have presented an optimization algorithm, with which the optimal maintenance interval can be determined from an availability interval point of view. A series of numerical experiments indicated that significant improvements in the availability interval can be obtained in comparison with a classical limiting availability perspective, and that these effects become stronger as the variabilities in life and/or repair times increase. Simply stated, our computational results have illustrated that preventive maintenance does not only increase the availability, but also reduces the

Table 4.2: Availability interval percentiles for T_u at the optimal maintenance intervals θ^* for interval availability and θ_0 for limiting availability, for 16 test problems with $u = 1$, $\mu_L = 1$, and different value of ω.

					$\omega = 0.90$			$\omega = 0.95$			$\omega = 0.99$		
$\frac{\mu_R}{\mu_L}$	$\frac{\mu_P}{\mu_R}$	$\frac{\sigma_L}{\mu_L}$	$\frac{\sigma_R}{\mu_R}$	θ_0	θ^*	$f_u^\omega(\theta^*)$	$f_u^\omega(\theta_0)$	θ^*	$f_u^\omega(\theta^*)$	$f_u^\omega(\theta_0)$	θ^*	$f_u^\omega(\theta^*)$	$f_u^\omega(\theta_0)$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1.10	0.37	1.93	2.05	0.40	2.13	2.27	0.40	2.52	2.71
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	1.10	0.64	1.89	1.96	0.64	2.03	2.11	0.43	2.30	2.39
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	0.78	0.50	1.50	1.71	0.50	1.50	1.89	0.50	1.74	2.26
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0.78	0.50	1.50	1.70	0.50	1.50	1.81	0.34	1.75	2.01
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	0.56	0.26	1.50	1.82	0.20	1.71	2.03	0.22	2.23	2.48
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0.56	0.26	1.50	1.76	0.37	1.84	1.91	0.28	2.07	2.25
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0.63	0.50	1.25	1.25	0.50	1.25	1.54	0.40	1.38	1.94
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0.63	0.50	1.25	1.25	0.50	1.25	1.56	0.40	1.38	1.77
$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1.10	0.36	1.46	1.51	0.39	1.56	1.62	0.40	1.76	1.84
$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	1.10	0.62	1.44	1.48	0.62	1.51	1.55	0.43	1.65	1.69
$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	0.78	0.50	1.25	1.35	0.50	1.25	1.44	0.50	1.36	1.62
$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0.78	0.50	1.25	1.35	0.50	1.25	1.40	0.50	1.37	1.51
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	0.56	0.26	1.25	1.40	0.20	1.32	1.50	0.21	1.61	1.72
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0.56	0.26	1.25	1.37	0.37	1.42	1.44	0.28	1.53	1.62
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0.63	0.50	1.13	1.13	0.50	1.13	1.26	0.39	1.19	1.46
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0.63	0.50	1.13	1.13	0.50	1.13	1.28	0.39	1.19	1.39

variability of a production system, and that the latter is often a more important performance measure. Although these conclusions were drawn within a setting of Gamma distributed repair and fixed maintenance times, we strongly believe that they are also applicable to more complex systems.

To conclude this chapter, let us now briefly discuss the possibilities for approximating the availability interval distribution, in case repair times are not Gamma distributed random variables and/or preventive maintenance times are not fixed. In the most general case, it not possible to derive explicit formulas for the convolutions $G_m \circ H_n$ appearing in $P_k(T_u \leq t)$. Under such circumstances, another interesting and potentially promising approach is to fit a Gamma (or other) distribution to the first two moments $E_\infty\{T_u - u\}$ and $E_\infty\{(T_u - u)^2\}$ of $T_u - u$. The underlying observation behind this approach is that these moments can be determined as long as the first three moments of $G(.)$ and $H(.)$ are available. In general, we may have to account for the fact that the availability interval distribution $P_\infty(T_u \leq t)$ might have some discontinuities as well. If these jumps are known in advance, i.e. in terms of a set $\Omega = \{t \geq u \mid P_\infty(T_u = t) > 0\}$ of availability intervals with non-zero probabilities, it seems worthwile to approximate $P_\infty(T_u \leq t \mid t \notin \Omega\}$ with the use of $E_\infty\{T_u - u \mid t \notin \Omega\}$ and $E_\infty\{(T_u - u)^2 \mid t \notin \Omega\}$. These suggestions, however, are left for future research.

4.9 Appendix

Proof of Lemma 5

As a starting point, we observe that the following expressions can be derived for the cumulative distribution functions $\Phi_3(.)$ and $\Phi_4(.)$ of the remaining preventive down time $R_3 \in [0, \mu_P]$, and the remaining corrective down time $R_4 \in [0, \infty\rangle$:

$$\Phi_3(\tau) = \frac{\tau}{\mu_P} \quad , \quad 0 \leq \tau \leq \mu_P$$

$$\Phi_4(\tau) = \frac{1}{\mu_R} \cdot \int_0^\tau 1 - \Gamma_{\alpha,\beta}(v)\, dv \quad , \quad \tau \geq 0$$

For notational convenience, and without loss of generality, we substitute $z = t - u - m \cdot \mu_P$ in the sequel. Our analysis now proceeds as follows. First of all, we observe that the expressions for $G_m \circ H_n(t-u)$ and $\Phi_3 \circ G_m \circ H_n(t-u)$ can be derived rather straightforwardly by some elementary algebra. On the other hand,

$\Phi_4 \circ G_m \circ H_n(t-u)$ yields a somewhat more complex expression, which must be further simplified:

$$G_m \circ H_n(t-u) = H_n(z) = \Gamma_{n\cdot\alpha,\beta}(z)$$

$$\begin{aligned}
\Phi_3 \circ G_m \circ H_n(t-u) &= \Phi_3 \circ \Gamma_{n\cdot\alpha,\beta}(z) \\
&= \frac{1}{\mu_P} \cdot \int_0^{\mu_P} \Gamma_{n\cdot\alpha,\beta}(z-\tau)\ d\tau \\
&= \frac{\Psi_{n\cdot\alpha,\beta}(z) - \Psi_{n\cdot\alpha,\beta}(z-\mu_P)}{\mu_P}
\end{aligned}$$

$$\begin{aligned}
\Phi_4 \circ G_m \circ H_n(t-u) &= \Phi_4 \circ \Gamma_{n\cdot\alpha,\beta}(z) \\
&= \frac{1}{\mu_R} \cdot \int_0^{z} (1 - \Gamma_{\alpha,\beta}(\tau)) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\tau)\ d\tau \\
&= \frac{\Psi_{n\cdot\alpha,\beta}(z) - \int_0^{z} \Gamma_{\alpha,\beta}(\tau) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\tau)\ d\tau}{\mu_R}
\end{aligned}$$

Apparently, we need to show that $\Psi_{(n+1)\cdot\alpha,\beta}(z) = \int_0^z \Gamma_{\alpha,\beta}(\tau) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\tau)\ d\tau$ in order to arrive at the expression for $\Phi_4 \circ G_m \circ H_n(t-u)$ in Lemma 5. Our analysis now proceeds as follows. First of all, we observe that $\int_0^z \Gamma_{\alpha,\beta}(\tau) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\tau)\ d\tau$ can be rewritten as follows:

$$\begin{aligned}
\int_0^z \Gamma_{\alpha,\beta}(\tau) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\tau)\ d\tau &= \int_0^z \int_0^\tau \frac{d}{d\upsilon} \left\{\Gamma_{\alpha,\beta}(\upsilon) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\upsilon)\right\}\ d\upsilon\ d\tau \\
&= \int_0^z \int_0^\tau \gamma_{\alpha,\beta}(\upsilon) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\upsilon)\ \ d\upsilon\ d\tau - \int_0^z \int_0^\tau \Gamma_{\alpha,\beta}(\upsilon) \cdot \gamma_{n\cdot\alpha,\beta}(z-\upsilon)\ \ d\upsilon\ d\tau
\end{aligned}$$

By changing the integration variables, both integrals can be reduced to one-dimensional integrals, each of which can be evaluated explicitly by using the following well-known properties $\int_0^z \tau \cdot \gamma_{\alpha,\beta}(\tau)\ d\tau = \alpha \cdot \beta \cdot \Gamma_{\alpha+1,\beta}(z)$ and $\int_0^z \gamma_{\alpha_1,\beta}(\tau) \cdot \Gamma_{\alpha_2,\beta}(z-\tau)\ d\tau = \Gamma_{\alpha_1+\alpha_2,\beta}(z)$ for Gamma distributions:

$$\int_0^z \int_0^\tau \gamma_{\alpha,\beta}(\upsilon) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\upsilon)\, d\upsilon\, d\tau = \int_0^z (z-\upsilon) \cdot \gamma_{\alpha,\beta}(\upsilon) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\upsilon)\, d\upsilon$$

$$= z \cdot \int_0^z \gamma_{\alpha,\beta}(\upsilon) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\upsilon)\, d\upsilon - \int_0^z \upsilon \cdot \gamma_{\alpha,\beta}(\upsilon) \cdot \Gamma_{n\cdot\alpha,\beta}(z-\upsilon)\, d\upsilon$$

$$= z \cdot \Gamma_{(n+1)\cdot\alpha,\beta}(z) - \alpha \cdot \beta \cdot \Gamma_{(n+1)\cdot\alpha+1,\beta}(z)$$

$$\int_0^z \int_0^\tau \Gamma_{\alpha,\beta}(\upsilon) \cdot \gamma_{n\cdot\alpha,\beta}(z-\upsilon)\, d\upsilon\, d\tau = \int_0^z (z-\upsilon) \cdot \Gamma_{\alpha,\beta}(\upsilon) \cdot \gamma_{n\cdot\alpha,\beta}(z-\upsilon)\, d\upsilon$$

$$= \int_0^z \upsilon \cdot \gamma_{n\cdot\alpha,\beta}(\upsilon) \cdot \Gamma_{\alpha,\beta}(z-\upsilon)\, d\upsilon = n \cdot \alpha \cdot \beta \cdot \Gamma_{(n+1)\cdot\alpha+1,\beta}(z)$$

Since $z \cdot \Gamma_{(n+1)\cdot\alpha,\beta}(z) - (n+1) \cdot \alpha \cdot \beta \cdot \Gamma_{(n+1)\cdot\alpha+1,\beta}(z) = \Psi_{(n+1)\cdot\alpha,\beta}(z)$, this completes the proof.

Chapter 5

Two-stage generalized age maintenance of an intermittently used production system

In general, the initiation of preventive maintenance should be based on the technical state as well as the operating state of a production system. Since the operating state of a production system is often subject to fluctuations in time, the planning of preventive maintenance at preset moments (e.g. age/block replacement) cannot be optimal. To avoid this, we consider a so-called two-stage maintenance policy, which - in a first stage - uses the technical state of the production system to determine a finite interval $[t, t + \Delta t]$ during which preventive maintenance must be carried out, and - in a second stage - uses the operating state of the production system to determine the optimal starting time for preventive maintenance within this interval. A generalized age maintenance policy optimizing both t and Δt is formulated in the first stage. To this end, the actual starting time of preventive maintenance is modelled in terms of a uniform distribution over the maintenance interval. Moreover, the expected costs of preventive maintenance are modelled as a decreasing function of the interval size. An efficient algorithm is developed to demonstrate the optimal maintenance strategies via numerical results that offer useful insights.

5.1 Introduction

Every few years, new surveys appear on maintenance optimization, showing the use and benefits of mathematical models in the maintenance area, e.g. see McCall (1965), Pierskalla and Voelker (1976), Sherif and Smith (1981), Valdez-Flores and Feldman (1989) and Cho and Parlar (1991). Most of these models, however, fail to incorporate special characteristics of the production facilities. Typically, this is the case

in commonly used mathematical models for maintenance optimization, such as the age and block replacement model (Barlow and Proschan 1965), the modified block replacement model (Berg and Epstein 1976), the minimal repair model (Barlow and Hunter 1960), the standard inspection model (Barlow, Hunter, and Proschan 1963), and the delay-time model (Christer and Waller 1984). Apart from the failure statistics, all information on the effects of down-time on the production system have to be condensed into two constants $c_p > 0$ and $c_f > 0$, representing the expected costs of preventive and corrective maintenance respectively.

In our view, the costs associated with preventive and corrective maintenance should be divided into direct maintenance costs (e.g. salaries of maintenance personnel, spare parts, tools) and indirect maintenance costs (e.g. production loss, delay penalties, holding costs). Simply stated, direct maintenance costs depend on the **technical state** of the production system (e.g. age of the machine, failure characteristics), and indirect maintenance costs on the **operating state** of the production system (e.g. buffer contents, workload, due dates). In general, the initiation of corrective maintenance is only based on the technical state of the production system. After all, most production systems are repaired at failure, disregarding the operating state of the production system.

On the other hand, the initiation of preventive maintenance should be based on the technical state as well as the operating state of the production system, in order to reduce the impact of indirect preventive maintenance costs. Since the operating state of a production system is often subject to fluctuations in time, there is a perspective of significant gains if some flexibility is built in concerning the starting time of preventive maintenance (e.g. in a setting of production orders with release and due dates). It seems reasonable therefore to consider a so-called two-stage maintenance policy, which - in a first stage - determines a finite interval $[t, t+\Delta t]$ during which preventive maintenance must be carried out, and - in a second stage - determines the optimal starting time $\hat{t}$ for preventive maintenance within this interval. Although this is a widespread common sense in practice, it certainly is an underexposed point of view in existing literature.

Simply stated, the design of our maintenance policy is based upon a hierarchical decomposition principle, where we distinguish between two planning stages (see Figure 5.1). In the first stage (the long term), information about the operating state of the production system is assumed to be available in terms of a stationary stochastic process only. Therefore, the initiation of preventive maintenance is based on the technical state of the production system. In the second stage (the short term), information about the operating state of the production system is assumed to be

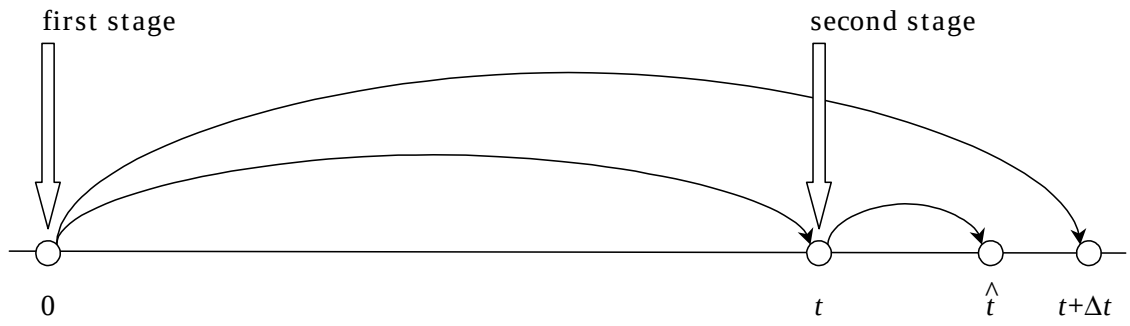


Figure 5.1: Decomposition principle of the two-stage maintenance policy, which - in a first stage - determines a finite interval $[t, t + \Delta t]$ during which preventive maintenance must be carried out, and - in a second stage - determines the optimal starting time $\hat{t}$ for preventive maintenance within that interval.

available beforehand over a finite, rolling horizon. Therefore, the initiation of preventive maintenance is based on the operating state of the production system. Of course, the crucial part of our approach will be to define adequate models for both stages, and to connect them properly. At least, first-stage models should incorporate second-stage implications, and vice versa.

5.1.1 Related literature

Although our two-stage maintenance concept can be observed in a variety of practical situations (a typical application was found at Vliegbasis Twenthe, one of the main operating bases of the Dutch Royal Airforce, where F16's undergo overhaul maintenance between 190 and 210 flight hours), we are not aware of any mathematical models that support this type of preventive maintenance planning, at least in existing literature. Nevertheless, several authors have recognized that the initiation of preventive maintenance should also be based on the operating state of a production system, and not on its technical state only. Here, we will only mention some important and illustrative references.

As a starting point, there is a considerable amount of literature which deals with the modelling and optimization of so-called opportunistic maintenance policies. The main underlying observation behind these models is that preventive maintenance activities can only be carried out at times when the system is not required, or not available for production. Sometimes, failure repairs of one or more components create opportunities for preventive maintenance on other components, e.g. see Jorgenson et al. (1967), Berg (1978), Bäckert and Rippin (1985), Van der Duyn Schouten and

Vanneste (1990), and Zheng (1995). It is also possible that maintenance opportunities are generated independently of component failures, e.g. by idle times in the production schedule, or by withdrawn production orders. Typical examples of such models can be found in e.g. Mine et al. (1981), Berg (1984), Dekker and Dijkstra (1992), Dekker and Smeitink (1994), and Dagpunar (1996). The reader is referred to Dekker and Smeitink (1991) for an extensive literature review on opportunistic (block) replacement models.

In the past few decades, much attention has been paid to the performance behavior of production/inventory systems with random service disruptions, e.g. see De Koster (1988), Lee and Rosenblatt (1987), Groenevelt et al. (1992), Berg et al. (1994), and Moinzadeh and Aggarwal (1997). A basic contribution is the paper of Wijngaard (1979), who considers a flow line consisting of two machines with exponentially distributed up and down times, which are connected by an intermediate buffer with finite storage capacity. Although the impact of machine failures on system performance is widely recognized, there are only few papers which explicitly account for preventive maintenance policies within the context of production/inventory systems, e.g. see Lee and Rosenblatt (1989), Van der Duyn Schouten and Vanneste (1995), Meller and Kim (1996), and Srinivasan and Lee (1996). Moreover, these models are often very complex, and only apply to specific settings.

In our opinion, there is a lack of elementary maintenance concepts and accompanying models, which more explicitly take into account that (i) the initiation of preventive maintenance should also be based on the operating state of a production system, and (ii) this operating state is often known in advance over a finite rolling horizon in an operational planning phase. The newly developed two-stage maintenance concept is our first attempt into this direction. Throughout this chapter, we will mainly focus on its applications into the classical age replacement policy (Barlow and Proschan 1965). Nevertheless, our approach could easily be generalized to other maintenance concepts as well (e.g. block replacement, minimal repair, and inspection models). We will come back to that later on in this chapter.

5.1.2 Outline

The outline of this chapter is as follows. In section 5.2, we present the general approach of our two-stage maintenance policy, and elaborate upon the interaction between both stages. In section 5.3, a generalized age maintenance policy optimizing both t and Δt is presented, in which the second stage is incorporated by means of an approximate model. In section 5.4, the second stage is studied in more detail for

an on/off production system, and a queue-like production system as well. Computational results in section 5.5 are given to validate our first stage model assumptions. Moreover, they indicate that significant savings can be obtained in comparison with a classical age maintenance policy. Finally, some concluding remarks are summarized in section 5.6.

5.2 General approach

Consider an unreliable production system, which is subject to random failures or breakdowns. In case of a failure, the system is repaired correctively. In addition, preventive maintenance actions are carried out to prevent failures. After each preventive and corrective maintenance action, the production system is assumed to be restored into an as-good-as-new condition. More specifically, the times between failures are mutually independent stochastic variables with cumulative distribution function $F(t)$ and corresponding probability density $f(t)$, where $t \geq 0$ denotes the elapsed (calendar or operating) time since the last maintenance action, either preventive or corrective.

As mentioned before, the design of our two-stage maintenance policy is based upon a hierarchical decomposition principle, where we distinguish between two planning levels. The highest level (the first stage) is related to the long term between preventive maintenance actions. The lowest level (the second stage) relates to a much shorter term, which is typical for the operational control of the production system. The main difference between these two levels is the kind of information which is available.

In the first stage, information about the operating state of the production system is assumed to be available in terms of a stationary stochastic process $X(\cdot)$ only. Therefore, the initiation of preventive maintenance is based upon the technical state (e.g. cumulative operating time) of the production system: preventive maintenance must be carried out somewhere between times t and $t + \Delta t$ since the last maintenance action, either preventive or corrective. In the second stage, information about the operating state of the production system, i.e. the actual realization of $X(\cdot)$, is assumed to be available beforehand over a finite, rolling horizon. Therefore, the initiation of preventive maintenance is based upon the operating state $X(\cdot)$ of the production system, which is now deterministically given over the entire maintenance interval $[t, t + \Delta t]$: preventive maintenance is carried out at the best opportunity $\hat{t}$ within this interval.

In general, our approach might invoke an upper bound Δt_{max} on Δt, since this kind of information is only available on a short term basis. For notational convenience, and without loss of generality, we assume that $\Delta t_{max} < \infty$ in the sequel. Now the

basic motivation for our approach is that both stages should interact in a simple but effective way, allowing separate control mechanisms for each stage. In line with this, our modelling framework proceeds as follows. In the first stage, a generalized age maintenance policy optimizing both t and Δt is formulated, in which the second stage is incorporated by means of an approximate model with two main elements:

- the actual starting time for preventive maintenance (as determined in the second stage) is modelled in terms of a uniform distribution over the maintenance interval $[t, t + \Delta t]$;
- the actual cost of preventive maintenance (as observed in the second stage) is modelled as an expected cost function $c_p(\Delta t)$, whereas c_f denotes the expected cost of corrective maintenance.

Although our analysis could easily be generalized to other than uniform distributions, we will restrict ourselves to the uniform distribution in this chapter. From a practical point of view, these assumptions relate to the uncertainty in operational information in the first stage, and are based on the (intuitive) reasoning that (i) each starting time within the interval $[t, t + \Delta t]$ has the same probability of being chosen, and (ii) there is no correlation between the actual starting time and the actual cost of preventive maintenance. In general, this reasoning can only hold approximately, since the second stage is controlled by complex stochastic processes, with their own stochastic laws (see section 5.4.2). There are situations, however, where these first stage assumptions are entirely justified from a second stage point of view (see section 5.4.1).

5.3 The first stage

In the first stage, information about the operating state of the production system is assumed to be available in terms of a stationary stochastic process $X(\cdot)$ only. Therefore, the initiation of preventive maintenance is driven by the technical state of the production system: preventive maintenance must be carried out somewhere between times t and $t + \Delta t$ since the last maintenance action, either preventive or corrective.

5.3.1 Model and assumptions

First of all, let us denote with $h(t, \Delta t)$ the expected costs of the two-stage maintenance policy per unit of time, and assume that analytical expressions for c_f and

$c_p(\Delta t)$ are available. Furthermore, define a cycle as the time between two consecutive maintenance actions, either preventive or corrective. Then it follows from renewal theory (Cox 1962) that the expected costs per unit of time are equal to the expected costs per cycle divided by the expected length of a cycle. Elaborating on the last two quantities yields the following expression for $h(t, \Delta t)$. Here, we denote $\overline{F}(t) \equiv 1 - F(t)$ for notational convenience:

$$h(t, \Delta t) = \frac{\int_t^{t+\Delta t} \left\{ c_f \cdot F(u) + c_p(\Delta t) \cdot \overline{F}(u) \right\} \, du}{\int_t^{t+\Delta t} \int_0^u \overline{F}(v) \, dv \, du}$$

Note that $g(t) = \lim_{\Delta t \to 0} h(t, \Delta t)$ coincides with the classical age maintenance policy (Barlow and Proschan 1965), which prescribes preventive maintenance to be carried out t time units after the last maintenance action, either preventive or corrective. Here, we denote $c_p = c_p(0)$ for notational convenience:

$$g(t) = \frac{c_f \cdot F(t) + c_p \cdot \overline{F}(t)}{\int_0^t \overline{F}(u) \, du}$$

Of course, our objective now is to minimize $h(t, \Delta t)$ with respect to t and Δt. Given that $F(t)$ and $c_p(\Delta t)$ are both **continuously differentiable** in t and Δt, it follows that $h(t, \Delta t)$ is also continuously differentiable in t and Δt. To simplify our analysis, we make the following (natural) assumptions with respect to the maintenance cost functions c_f and $c_p(\Delta t)$, as well as the failure rate $r(t) = f(t)/\overline{F}(t)$. Here, $\mu = \int_0^\infty \overline{F}(t) \, dt$ denotes the expected life time of the production system:

(i) $c_p(0) < c_f$,

(ii) $c_p(\Delta t)$ is decreasing in Δt,

(iii) $r(t)$ is strictly increasing in t,

(iv) $\lim_{t \to \infty} r(t) > \frac{c_f}{c_f - c_p(0)} \cdot \frac{1}{\mu}$.

In general, finding the optimal two-stage maintenance policy $\{t^*, \Delta t^*\}$ is a complex problem, since $h(t, \Delta t)$ cannot be evaluated explicitly, and may have multiple local minima as well. Therefore, our first objective in this section is to derive several properties of $h(t, \Delta t)$ with which optimal policies $\{t^*, \Delta t^*\}$ can be classified. With these properties, an efficient algorithm is developed to determine the optimal two-stage maintenance policy.

5.3.2 Generalized age maintenance

As a starting point of our analysis, we will show that the two-stage maintenance policy is indeed a generalized age maintenance policy. In other words, we need to prove that $\Delta t^* = 0$ if the costs of preventive maintenance $c_p(\Delta t)$ are fixed, i.e. independent of Δt. To this end, let us denote with t^* and Δt^* values of t and Δt for which $h(t, \Delta t)$ is minimized, and define $r(t, \Delta t)$ as follows:

$$r(t, \Delta t) = \frac{\int_t^{t+\Delta t} f(u)\ du}{\int_t^{t+\Delta t} \overline{F}(u)\ du}$$

From a practical point of view, $r(t, \Delta t)$ could be interpreted as a sort of averaged failure rate during the interval $[t, t + \Delta t]$. Now putting the derivatives of $h(t, \Delta t)$ with respect to t and Δt equal to zero, yields the following **equibrilium equations** for $h(t^*, \Delta t^*)$:

$$h(t^*, \Delta t^*) = (c_f - c_p(\Delta t^*)) \cdot r(t^*, \Delta t^*)$$

$$h(t^*, \Delta t^*) = \frac{c_f \cdot F(t^* + \Delta t^*) + c_p(\Delta t^*) \cdot \overline{F}(t^* + \Delta t^*) + c_p'(\Delta t^*) \cdot \int_{t^*}^{t^*+\Delta t^*} \overline{F}(u)\ du}{\int_0^{t^*+\Delta t^*} \overline{F}(u)\ du}$$

If $\Delta t^* < \Delta t_{max}$, these equations represent sufficient conditions for a local extremum $\{t^*, \Delta t^*\}$, and necessary conditions for a global minimum $\{t^*, \Delta t^*\}$. Moreover, if $\Delta t^* = 0$, they reduce to some well-known relations for the classical age maintenance policy. Here, we denote $c_p = c_p(0) < c_f$ for notational convenience:

$$g(t^*) = (c_f - c_p) \cdot r(t^*)$$

$$g(t^*) = \frac{c_f \cdot F(t^*) + c_p \cdot \overline{F}(t^*)}{\int_0^{t^*} \overline{F}(u)\ du}$$

Our analysis now proceeds as follows. First of all, we will show that $r(t, \Delta t)$ is strictly increasing in t and Δt, provided that $r(t)$ is strictly increasing in t. For the proof, we need the following lemma:

Lemma 6 *Consider two strictly positive and continuous functions $m(t)$ and $n(t)$, where $t \geq 0$. If $m(t)/n(t)$ is strictly increasing in t, then the following relations hold:*

(a) $m(a)/n(a) < \int_a^b m(u)\,du \;/\; \int_a^b n(u)\,du < m(b)/n(b)$ *for all* $a, b \geq 0$ *with* $a < b$,

(b) $\int_a^{a+b} m(u)\,du \;/\; \int_a^{a+b} n(u)\,du$ *is strictly increasing in* $a \geq 0$ *for all* $b \geq 0$,

(c) $\int_a^{a+b} m(u)\,du \;/\; \int_a^{a+b} n(u)\,du$ *is strictly increasing in* $b \geq 0$ *for all* $a \geq 0$.

Proof. See Appendix A. □

Theorem 1 *$r(t, \Delta t)$ is strictly increasing in $t \geq 0$ for all $\Delta t \geq 0$, and strictly increasing in $\Delta t \geq 0$ for all $t \geq 0$; moreover, $r(t) < r(t, \Delta t) < r(t + \Delta t)$ for all $t \geq 0$ and $\Delta t > 0$.*

Proof. Since $r(t) = f(t)/\overline{F}(t)$ is strictly increasing in t by assumption (iii), substitution of $m(u) = f(u)$, $n(u) = \overline{F}(u)$, $a = t$ and $b = \Delta t$ in Lemma 6b yields: $r(t, \Delta t)$ is strictly increasing in $t \geq 0$ for all $\Delta t \geq 0$. In a similar way, substitution of $m(u) = f(u)$, $n(u) = \overline{F}(u)$, $a = t$ and $b = \Delta t$ in Lemma 6c yields: $r(t, \Delta t)$ is strictly increasing in $\Delta t \geq 0$ for all $t \geq 0$. Finally, it follows from Lemma 6a that $r(t) < r(t, \Delta t) < r(t + \Delta t)$ for all $t \geq 0$ and $\Delta t > 0$, which completes the proof. □

With these properties of $r(t, \Delta t)$ in mind, it can now easily be derived that $\Delta t^* = 0$ if the costs of preventive maintenance $c_p(\Delta t)$ are independent of Δt. This is stated more explicitly in the following theorem.

Theorem 2 *If $c_p(\Delta t) = c_p > 0$ for all $\Delta t \geq 0$, then $h(t, \Delta t)$ has a unique minimum $\{t^*, \Delta t^*\} = \{t_0, 0\}$, where $t_0 < \infty$ reflects the optimal age maintenance policy. Moreover, $min\{h(t, \Delta t) \mid t \geq 0\}$ is strictly increasing in $\Delta t \geq 0$.*

Proof. As a result of assumptions (iii) and (iv), $g(t) = \lim_{\Delta t \to 0} h(t, \Delta t)$ has a unique minimum $t_0 < \infty$ with $g(t_0) = (c_f - c_p) \cdot r(t_0)$, e.g. see Barlow and Proschan (1965) for further details. As a consequence, $g(t) > g(t_0)$ for all $t \neq t_0$. Let us now consider an optimal two-stage maintenance policy $\{t^*, \Delta t^*\}$. With $c_p(\Delta t) = c_p > 0$, and thus $c_p'(\Delta t) = 0$ for all $\Delta t \geq 0$, the equibrilium equations for $h(t^*, \Delta t^*)$ reduce to:

$$h(t^*, \Delta t^*) = (c_f - c_p) \cdot r(t^*, \Delta t^*) = g(t^* + \Delta t^*)$$

Since $h(t^*, \Delta t^*) \leq g(t_0)$ by definition, and $g(t) > g(t_0)$ for all $t \neq t_0$, we have $t^* + \Delta t^* = t_0$. Now suppose that $\Delta t^* > 0$. Since $r(t^*) < r(t^*, \Delta t^*) < r(t^* + \Delta t^*)$ by Theorem 1, this yields $h(t^*, \Delta t^*) = (c_f - c_p) \cdot r(t^*, \Delta t^*) < (c_f - c_p) \cdot r(t^* + \Delta t^*) = (c_f - c_p) \cdot r(t_0) = g(t_0) = g(t^* + \Delta t^*)$, which is in contradiction with $h(t^*, \Delta t^*) = g(t^* + \Delta t^*)$. Apparently, $\Delta t^* > 0$ cannot be optimal, and thus $\Delta t^* = 0$ and $t^* = t_0$ reflect the global minimum. Following a similar argument, it can be shown that $\min\{h(t, \Delta t) \mid t \geq 0\}$ is strictly increasing in $\Delta t \geq 0$. □

5.3.3 Problem decomposition

To continue our analysis, we will show that for fixed values of Δt, the minimization of $h(t, \Delta t)$ with respect to t is a relatively simple problem. As a starting point, we denote with $\xi(\Delta t) = \arg\min\{h(t, \Delta t) \mid t \geq 0\}$ the value of t for which $h(t, \Delta t)$ is minimized, and observe that the following implicit relation can be derived for $\xi(\Delta t)$:

$$(c_f - c_p(\Delta t)) \cdot r(\xi(\Delta t), \Delta t) = h(\xi(\Delta t), \Delta t)$$

Obviously, this equation reflects a necessary condition for $\xi(\Delta t)$, since each global minimum must coincide with a local extremum. In the following theorem, however, we will show that it is also a sufficient condition for $\xi(\Delta t)$. In other words, we will show that there exists exactly one local minimum $\xi(\Delta t)$, and no local maxima, for each $\Delta t \geq 0$. For the proof, we need the following lemma:

Lemma 7 *For each $\Delta t \geq 0$, t^* is a local minimum of $h(t, \Delta t)$ with respect to t, if and only if there exists an $\varepsilon > 0$ such that:*

$$(c_f - c_p(\Delta t)) \cdot r(t, \Delta t) - h(t, \Delta t) \begin{cases} < 0 & for \quad t^* - \varepsilon < t < t^* \\ = 0 & for \quad t = t^* \\ > 0 & for \quad t^* < t < t^* + \varepsilon \end{cases}$$

Proof. First of all, recall that t^* is a local minimum if and only if there exists an $\varepsilon > 0$ such that $\frac{\partial h}{\partial t} \leq 0$ for $t^* - \varepsilon < t \leq t^*$, and $\frac{\partial h}{\partial t} \geq 0$ for $t^* \leq t < t^* + \varepsilon$. Now suppose that $\frac{\partial h}{\partial t} = 0$ on a finite interval $[t_1, t_2]$ with $t_1 \leq t^* \leq t_2$ and $t_1 < t_2$. Then t_1 and t_2 are also local extrema, and thus $h(t_1, \Delta t) = (c_f - c_p(\Delta t)) \cdot r(t_1, \Delta t)$ and $h(t_2, \Delta t) = (c_f - c_p(\Delta t)) \cdot r(t_2, \Delta t)$. Since $r(t, \Delta t)$ is strictly increasing in t by Theorem 1, and $t_1 < t_2$ by construction, this yields $h(t_1, \Delta t) < h(t_2, \Delta t)$, which is in conflict with the assumption that $\frac{\partial h}{\partial t} = 0$ on $[t_1, t_2]$. Apparently, $t_1 < t_2$ leads to a contradiction. Hence, $t_1 = t^* = t_2$, implying that $\frac{\partial h}{\partial t} < 0$ for $t^* - \varepsilon < t < t^*$, $\frac{\partial h}{\partial t} = 0$ for $t = t^*$, and $\frac{\partial h}{\partial t} > 0$ for $t^* < t < t^* + \varepsilon$. □

Theorem 3 *There exists a unique local minimum $\xi(\Delta t) < \infty$; moreover,*

$$(c_f - c_p(\Delta t)) \cdot r(t, \Delta t) - h(t, \Delta t) \begin{cases} < 0 & for \quad t < \xi(\Delta t) \\ = 0 & for \quad t = \xi(\Delta t) \\ > 0 & for \quad t > \xi(\Delta t) \end{cases}$$

Proof. As a starting point, we observe that $h(\xi(\Delta t), \Delta t) \leq c_f \cdot \mu^{-1}$, since we can never do worse than a purely corrective maintenance strategy. Since $\lim_{t\to\infty}(c_f - c_p(\Delta t)) \cdot r(t, \Delta t) > (c_f - c_p(0)) \cdot \lim_{t\to\infty} r(t) > c_f \cdot \mu^{-1}$ by assumptions (ii)-(iv), it follows that $\xi(\Delta t) < \infty$ for all $\Delta t \geq 0$. Now suppose that t_1 and t_2 are both local minima of $h(t, \Delta t)$ given Δt, with $t_1 < t_2$. Since $h(t, \Delta t)$ is continuously differentiable in t, it follows from Lemma 7 that there must also exist a local maximum t^* with $t_1 < t^* < t_2$. In a similar way, this implies the existence of an $\varepsilon > 0$ such that $(c_f - c_p(\Delta t)) \cdot r(t, \Delta t) > h(t, \Delta t)$ for $t^* - \varepsilon < t < t^*$, and $(c_f - c_p(\Delta t)) \cdot r(t, \Delta t) < h(t, \Delta t)$ for $t^* < t < t^* + \varepsilon$. With $\varepsilon \to 0$, this requires $(c_f - c_p(\Delta t)) \cdot r(t, \Delta t)$, and thus $r(t, \Delta t)$, to be decreasing in $t = t^*$, which is in conflict with Theorem 1. Apparently, the assumption of two (or more) local minima leads to a contradiction. Hence, there is exactly one local minimum $\xi(\Delta t) < \infty$, and there are no local maxima. Consequently, $(c_f - c_p(\Delta t)) \cdot r(t, \Delta t) < h(t, \Delta t)$ for $t < \xi(\Delta t)$, and $(c_f - c_p(\Delta t)) \cdot r(t, \Delta t) > h(t, \Delta t)$ for $t > \xi(\Delta t)$. This implies the uniqueness of $\xi(\Delta t)$, and completes the proof. □ .

As a result of Theorem 3, we know now that $h(t, \Delta t)$ is a unimodal function in t for all $\Delta t \geq 0$ (a local minimum of a unimodal function is also a global minimum). Hence, $\xi(\Delta t)$ can be determined efficiently with the use of standard search techniques. Now, a natural idea is that a search for the global minimum $\{t^*, \Delta t^*\}$ of $h(t, \Delta t)$ should exploit this fact. Therefore, we have chosen to decompose our global optimization problem into two subproblems, one that determines $\xi(\Delta t)$ for a given value of Δt, and one that minimizes $h(\xi(\Delta t), \Delta t)$ with respect to Δt. This is stated more explicitly in the following section.

5.3.4 A branch and bound algorithm

In general, finding the optimal two-stage maintenance policy $\{t^*, \Delta t^*\}$ is a complex problem, since $h(t, \Delta t)$ cannot be evaluated explicitly, and may have multiple local minima as well. As a consequence, classical optimization procedures may get stuck in a local minimum, which is not desirable. To avoid this, we have developed an efficient numerical optimization algorithm, which combines the concept of bisection search

with branch and bound enumeration. This algorithm is based on the decomposition principle of the previous section, and proceeds as follows.

As a starting point, the optimal age maintenance policy $t_0 < \infty$ is determined, resulting in a best-so-far two-stage maintenance policy $\{t^*, \Delta t^*\} = \{t_0, 0\}$, with corresponding costs $g(t_0)$. Subsequently, a node $(0, \Delta t_{max})$ is created, where $\Delta t_{max} < \infty$ denotes the maximal interval size. In general, each node (a, b) corresponds with an interval $a \leq \Delta t \leq b$, and contains a lower bound $h_{low}(a, b)$ for the optimal policy within that interval. If $h_{low}(a, b) \cdot (1 + \varepsilon) > h(t^*, \Delta t^*)$, where $\{t^*, \Delta t^*\}$ denotes the best-so-far policy and $\varepsilon > 0$ is a user-defined constant, the corresponding node (a, b) is closed. Otherwise, a closer look at that interval is necessary, and bisection is used. More specifically, new nodes (a, c) and (c, b) with $c = \frac{a+b}{2}$ are created, and $\xi(c)$ is determined to check whether the best-so-far policy $\{t^*, \Delta t^*\}$ is outperformed by $\{\xi(c), c\}$. Subsequently, $h_{low}(a, c)$ and $h_{low}(c, b)$ are determined, and the most promising node - in terms of the corresponding lower bound - is selected. This procedure is repeated recursively until all nodes are closed. At this point, $\{t^*, \Delta t^*\}$ denotes a so-called ε-optimal two-stage maintenance policy.

The question remains how to determine a lower bound $h_{low}(a, b)$ for the optimal two-stage maintenance policy within the range $a \leq \Delta t \leq b$. First of all, we observe that $c_p(\Delta t)$ is decreasing in Δt, and thus $c_p(\Delta t) \geq c_p(b)$ for all $\Delta t \leq b$. Moreover, if we substitute $c_p(\Delta t) = c_p(b)$ in $h(t, \Delta t)$, it follows from Theorem 2 that $\min\{h(t, \Delta t) \mid t \geq 0\} \geq \min\{h(t, a) \mid t \geq 0\}$ for all $\Delta t \geq a$. Summarizing, this leaves us with the following lower bound $h_{low}(a, b)$ for $\min\{h(t, \Delta t) \mid t \geq 0, a \leq \Delta t \leq b\}$:

$$h_{low}(a, b) = \min_{t \geq 0} \left\{ \frac{\int_t^{t+a} \left\{ c_f \cdot F(u) + c_p(b) \cdot \overline{F}(u) \right\} \, du}{\int_t^{t+a} \int_0^u \overline{F}(v) \, dv \, du} \right\}$$

Similar to the proof of Theorem 3, it can be shown that the right hand side of this equation is a unimodal function in t. In other words, $h_{low}(a, b)$ can be determined efficiently with the use of standard search techniques, in particular if a good initial value for t is provided. In our optimization algorithm, $h_{low}(a, c)$ and $h_{low}(c, b)$ are determined with an initial value $\xi(c)$ for t. In a similar way, $\xi(c)$ is determined through evaluation of $h_{low}(c, c)$, with an initial value for t that is inherited from the parental node. Further details are skipped, since they are not so relevant for what follows.

In general, $h(t, \Delta t)$ cannot be evaluated explicitly, as a result of which all integrals in t and Δt must be numerically approximated, for example by means of a trapezoidal

Table 5.1: Effect of $\delta > 0$ and $\epsilon > 0$ on the accuracy of the optimal policy $\{t^*, \Delta t^*\}$, for a production system with Gamma distributed life time: $\mu = \sigma^2 = 10$, $c_f = 100$, and $c_p(\Delta t) = 20 + 5 \cdot e^{-0.1 \cdot \Delta t}$.

δ	ε	t^*	Δt^*	$h(t^*, \Delta t^*)$	# seconds	# nodes
0.1	0.01	4.9	1.9	5.21841061	0.002	11
0.1	0.001	5.0	1.7	5.21686204	0.004	33
0.01	0.001	5.01	1.65	5.21629661	0.012	47
0.01	0.0001	5.02	1.64	5.21628726	0.058	125
0.001	0.0001	5.022	1.637	5.21628262	0.157	177
0.001	0.00001	5.021	1.638	5.21628260	0.238	411
0.0001	0.00001	5.0214	1.6377	5.21628255	0.583	555
0.0001	0.000001	5.0214	1.6376	5.21628255	0.797	1275

rule with step sizes δ, where $\delta > 0$ is a user-defined constant. As a consequence, only maintenance policies of the form $\{m \cdot \delta, n \cdot \delta\}$, with $m, n \in \mathbb{N}$, are considered. Obviously, smaller values of δ and ε lead to more accurate results, but also require more computational effort (see Table 5.1). In the remainder of this chapter, we will restrict ourselves to Gamma distributed life times $F(t) = \Gamma_{\alpha,\beta}(t)$, with mean $\mu = \alpha \cdot \beta$, and variance $\sigma^2 = \alpha \cdot \beta^2$. By doing this, $g(t)$ and $h(t, \Delta t)$ can be evaluated explicitly as follows. Once again, we denote $c_p = c_p(0)$ for notational convenience:

$$g(t) = \frac{c_p + (c_f - c_p) \cdot \Gamma_{\alpha,\beta}(t)}{t \cdot (1 - \Gamma_{\alpha,\beta}(t)) - \alpha \cdot \beta \cdot \Gamma_{\alpha+1,\beta}(t)}$$

$$h(t, \Delta t) = \frac{c_p(\Delta t) \cdot \Delta t + (c_f - c_p(\Delta t)) \cdot \left[u \cdot \Gamma_{\alpha,\beta}(u) - \alpha \cdot \beta \cdot \Gamma_{\alpha+1,\beta}(u)\right]_t^{t+\Delta t}}{\left[\frac{1}{2}u^2 \cdot (1 - \Gamma_{\alpha,\beta}(u)) + \alpha \cdot \beta \cdot u \cdot \Gamma_{\alpha+1,\beta}(u) - \frac{1}{2}(\alpha + \alpha^2) \cdot \beta^2 \cdot \Gamma_{\alpha+2,\beta}(u)\right]_t^{t+\Delta t}}$$

For the proof, we refer to Appendix B. Here, we only mention that efficient computer programming codes are available for the calculation of Gamma distributions, e.g. see Temme (1994). Of course, these procedures were further exploited in deriving the optimal two-stage maintenance policy $\{t^*, \Delta t^*\}$.

5.4 The second stage

In the second stage, information about the operating state of the production system, i.e. the actual realization of $X(\cdot)$, is assumed to be available beforehand over a finite, rolling horizon. Therefore, the initiation of preventive maintenance is based upon the

operating state $X(\cdot)$ of the production system: preventive maintenance is carried out at the best opportunity $\hat{t}$ between times t and $t + \Delta t$. In general, we have to account for the ensemble of all possible realizations of the process $X(\cdot)$ in the second stage, in order to arrive at the averages relevant for c_f and $c_p(\Delta t)$ in our first stage model. In this section, we will restrict ourselves to the case where $X(.)$ behaves according to (i) an on/off process with exponentially distributed on- and off-periods, and (ii) a continuous-time Markov process on a finite state space. To simplify our analysis, we assume that $X(\cdot)$ is ergodic (i.e. all states are recurrent), and that $X(\cdot)$ is restored to a stationary state after each preventive and corrective maintenance action. In the remainder of this section, these assumptions will be studied in more detail for both examples.

5.4.1 An on/off production system

As a starting point, we consider a production system with alternating on- and off-periods. During an on-period, the system produces continuously, and is subject to failure (the busy state). During an off-period, the production system is not productive, and cannot fail (the idle state). Our objective is to minimize the limiting conditional pointwise unavailability of the production system, i.e. the long term probability that the system is not available when needed for production. Therefore, preventive maintenance is planned at the largest off-period during t and $t + \Delta t$ units of cumulative operating time (see Figure 5.2). In addition, corrective maintenance actions are carried out upon failure.

To a certain extent, this approach is similar to the modelling framework suggested by Berg (1984), in which the system is maintained preventively as soon as it reaches age $d > 0$, or at the beginning of the first off-period at which its age exceeds $w \leq d$. Nevertheless, our modelling framework is much stronger, since it chooses the largest rather than the first off-period between ages w and d. If $w = d$, both policies coincide with a classical age maintenance policy.

Our analysis now proceeds as follows. First of all, the time required for preventive and corrective maintenance are modelled as random variables, with cumulative probability functions $G_p(\cdot)$ and $G_f(\cdot)$ respectively. For notational convenience, we restrict ourselves to the case where both on- and off-periods are mutually independent, exponentially distributed random variables, with means $\lambda^{-1}$and μ^{-1} respectively. Our analysis, however, could also be generalized to other than exponential distributions.

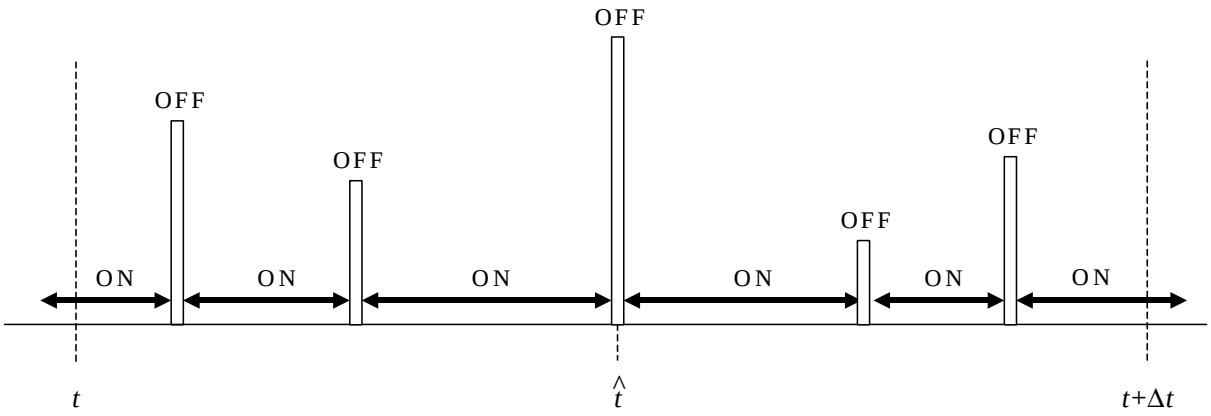


Figure 5.2: Preventive maintenance is planned at the largest off period.

The expected maintenance costs c_f and $c_p(\Delta t)$

As a starting point of our analysis, let us denote with $N_{\Delta t}$ the number of off-periods within an arbitrary interval of Δt units of cumulative operating time. Since the length of each on-period is exponentially distributed with mean λ^{-1}, it is obvious that $N_{\Delta t}$ follows a Poisson distribution with parameter $\lambda \cdot \Delta t$. For $n \geq 0$, this yields:

$$P(N_{\Delta t} = n) = \frac{(\lambda \cdot \Delta t)^n \cdot e^{-\lambda \cdot \Delta t}}{n!}$$

Now let $Y_{\Delta t}$ denote the length of the largest off-period during an arbitrary interval of Δt units of cumulative operating time. Given that $N_{\Delta t} = n$, $Y_{\Delta t}$ will take a value less than $y \geq 0$ if and only if each of the n independent off-periods takes a value less than y. Hence, the conditional probability $P(Y_{\Delta t} \leq y \mid N_{\Delta t} = n)$ is given by:

$$P(Y_{\Delta t} \leq y \mid N_{\Delta t} = n) = (1 - e^{-\mu \cdot y})^n$$

Now define $H_{\Delta t}(y) = P(Y_{\Delta t} \leq y)$ as the probability that the largest off-period during an arbitrary interval of Δt units of cumulative operating time does not exceed y units of time. With the definitions of $N_{\Delta t}$ and $Y_{\Delta t}$ in mind, this yields the following expression for $H_{\Delta t}(y)$:

$$H_{\Delta t}(y) = \sum_{n=0}^{\infty} P(Y_{\Delta t} \leq y, N_{\Delta t} = n) = \sum_{n=0}^{\infty} P(Y_{\Delta t} \leq y \mid N_{\Delta t} = n) \cdot P(N_{\Delta t} = n)$$

$$= \sum_{n=0}^{\infty} (1 - e^{-\mu \cdot y})^n \cdot \frac{(\lambda \cdot \Delta t)^n e^{-\lambda \cdot \Delta t}}{n!} = e^{-\lambda \cdot \Delta t} \cdot e^{\lambda \cdot \Delta t \cdot (1 - e^{-\mu \cdot y})} = e^{-\lambda \cdot \Delta t \cdot e^{-\mu \cdot y}}$$

First of all, we observe that $H_0(y) = 1$ for all $y \geq 0$, implying that the largest off-period during an infinitely small interval equals zero with probability 1. This is

consistent with our intuition. In a similar way, $H_{\Delta t}(0) = e^{-\lambda \cdot \Delta t}$ implies that the largest off-period during an arbitrary interval of size Δt equals zero with probability $e^{-\lambda \cdot \Delta t}$. Indeed, this corresponds to the probability that the length of a single on-period exceeds the amount of Δt units of time. For notational convenience, let $h_{\Delta t}(y)$ denote the corresponding probability density function, and define $\overline{H}_{\Delta t}(y) \equiv 1 - H_{\Delta t}(y)$:

$$h_{\Delta t}(y) = \lambda \cdot \mu \cdot \Delta t \cdot e^{-\mu \cdot y - \lambda \cdot \Delta t \cdot e^{-\mu \cdot y}}$$

Since failures can only occur during production (on-periods), the expected unavailability time in case of a failure equals the expected corrective maintenance time. This leaves us with the following expression for c_f:

$$c_f = \int_0^\infty x \; dG_f(x)$$

In an analogous way, the following expression can be derived for $c_p(\Delta t)$, i.e. the expected unavailililibity time associated with preventive maintenance at the largest off-period during an arbitrary interval of Δt units of cumulative operating time:

$$c_p(\Delta t) = \int_0^\infty \left\{ x \cdot H_{\Delta t}(0) + \int_0^x (x - y) \cdot h_{\Delta t}(y) \; dy + 0 \cdot \overline{H}_{\Delta t}(x) \right\} \; dG_p(x)$$

The first term within curly brackets refers to the situation where no off-periods occur during the entire maintenance interval. In the second term, the largest off-period reduces the unavailability time to a smaller, but still positive value. The third term describes the (zero) contribution due to the occurence of an off-period longer than the time required for preventive maintenance. If we restrict ourselves to constant preventive and corrective maintenance times $t_p < t_f$, these expressions for c_f and $c_p(\Delta t)$ can be further reduced to:

$$c_f = t_f$$

$$c_p(\Delta t) = t_p \cdot e^{-\lambda \cdot \Delta t} + \lambda \cdot \mu \cdot \Delta t \cdot \int_0^{t_p} (t_p - y) \cdot e^{-\mu \cdot y - \lambda \cdot \Delta t \cdot e^{-\mu \cdot y}} \; dy$$

After some elementary algebra, it can be shown that $c_p(0) = t_p < t_f = c_f$, $c_p(\Delta t)$ is decreasing in Δt, and $\lim_{\Delta t \to \infty} c_p(\Delta t) = 0$. Moreover, it is at least intuitively clear that preventive maintenance costs $c_p(\Delta t)$ will be reduced if λ increases (shorter on-periods) or μ decreases (longer off-periods). On the other hand, if λ and μ increase

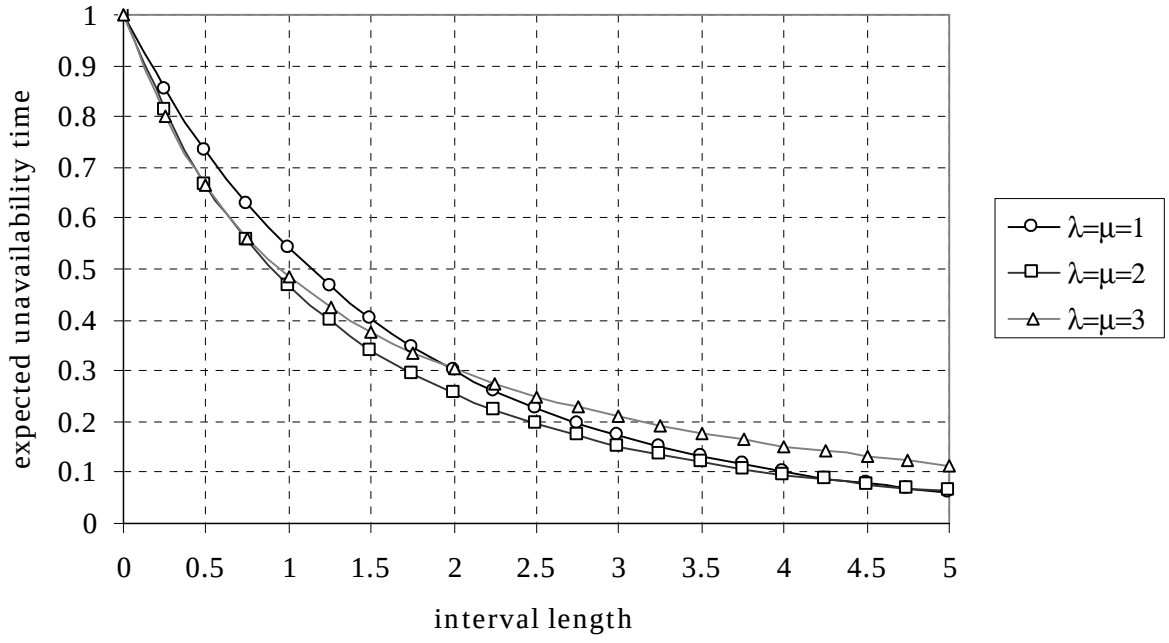


Figure 5.3: Expected unavailability time $c_p(\Delta t)$ as a function of the interval length Δt, for an on/off production system with constant preventive maintenance times $t_p = 1$, and different values for λ and μ.

or decrease with the same factor, and thus the occupation rate $\rho = \frac{\mu}{\lambda+\mu}$ of the production system remains unchanged, some interesting behavior can be observed. A typical example of this behavior is depicted in Figure 5.3. If Δt tends too zero, large values of λ and μ (i.e. frequent and short service interruptions) are to be preferred above small values of λ and μ (i.e. infrequent and long service interruptions). For large values of Δt, the opposite seems to be true.

Validation of the first stage model

Let us now examine whether the various assumptions that were made in our first stage model are valid. As a starting point, it follows from the complete randomness of Poisson processes (Heyman and Sobel 1982), that the location of the largest off-period is uniformly distributed over $[t, t + \Delta t]$, and that its length is independent of its location. For similar reasons, we may as well assume that the system is restored into a stationary state after each preventive and corrective maintenance action, since the length of each on-period is an exponentially distributed random variable. In other words, our first stage modelling assumptions are completely satisfied within

this setting of exponentially distributed on- and off-periods. Nevertheless, this reasoning would also apply to other than exponentially distributed off-periods. On the other hand, the assumption of exponentially distributed on-periods is essential for the validation of our first stage modelling assumptions.

5.4.2 A queue-like production system

Let us now consider the case where the operating state $X(\cdot)$ of the production system behaves according to a homogeneous continuous time Markov process on a finite discrete state space, without loss of generality denoted as $\{0, ..., m\}$. Transitions occur from state i to state $j \neq i$ with rate $q_{ij} \geq 0$. The system can be maintained at failure against expected cost $k_f(i)$, and preventively against expected cost $k_p(i) < k_f(i)$, where i denotes the operating state of the production system at the time maintenance is carried out. Without loss of generality, we assume that $k_p(0) < ... < k_p(m)$ and $k_f(0) < ... < k_f(m)$. Note that $k_p(\cdot)$ and $k_f(\cdot)$ may include direct as well as indirect maintenance costs. A simple example of this type is a service station with Poisson arrivals and exponentially distributed service times, in which indirect costs are incurred due to extra waiting, causing delays in delivery. Here, $m \geq 0$ denotes the storage capacity and $0 \leq i \leq m$ the number of jobs in the system.

The expected maintenance costs c_f and $c_p(\Delta t)$

Without loss of generality, we assume that $t = 0$ represents the start of the maintenance interval, and introduce $X_t \equiv X(t)$ for notational convenience. As a starting point of our analysis, we denote with $\mathbf{Q} = (q_{ij})$ the infinitesimal generator matrix of $X(\cdot)$, where $q_{ii} = -\sum_{j \neq i} q_{ij}$. Moreover, we let $Y_{\Delta t} = \min\{X_t \mid 0 \leq t \leq \Delta t\}$ denote the operating state $X_{\hat{t}}$ of the production system at the best opportunity $\hat{t}$ during the interval $[0, \Delta t]$, and define $p_{ijk}(t)$ as follows:

$$p_{ijk}(t) = P(Y_t = j, X_t = k \mid X_0 = i)$$

In order to obtain these transition probabilities, it is sufficient to define a Markov process $\{Z_t, t \geq 0\}$ as follows: $\{Z_t = (i, j)\} \sim \{Y_t = i, X_t = j\}$. Obviously, this yields the following alternative expression for $p_{ijk}(t)$:

$$p_{ijk}(t) = P(Z_t = (j, k) \mid Z_0 = (i, i))$$

It is now clear that $Z(t)$ is a homogeneous continuous time Markov process on a finite state space $\{(i, j) \mid 0 \leq i \leq j \leq m\}$, as is shown more explicitly in Figure

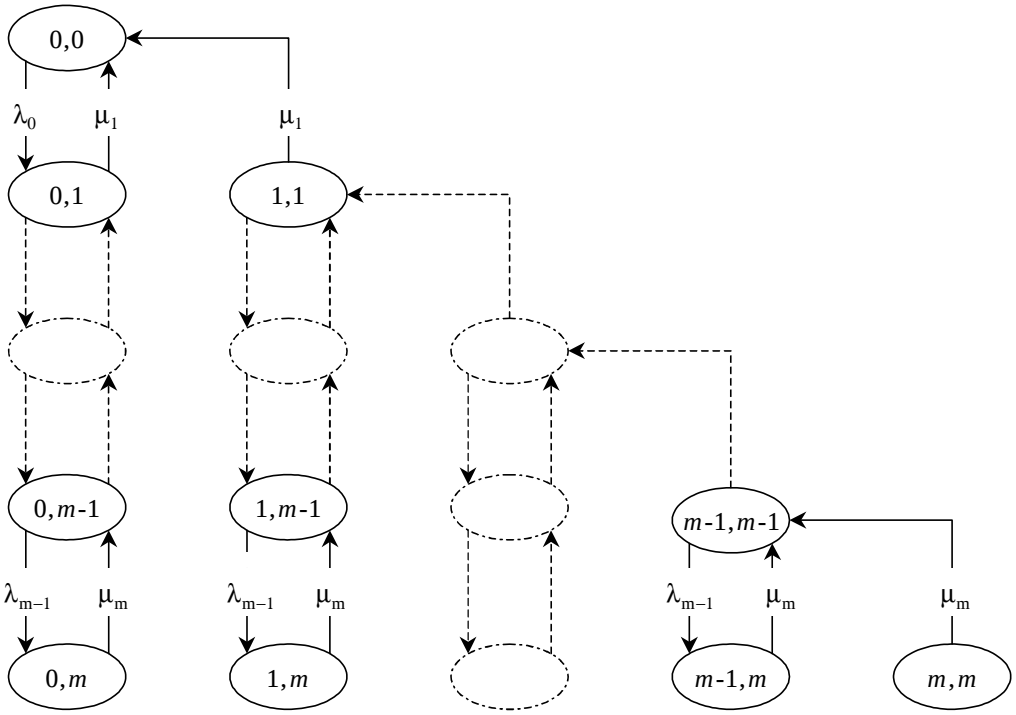


Figure 5.4: Transition diagram of $Z(.)$ in case $X(.)$ is a birth and death process on state space $\{0, ..., m\}$, with transition rates λ_i $(0 \leq i < m)$ and μ_i $(0 < i \leq m)$.

5.4. Hence, the conditional probabilities $p_{ijk}(t)$ can be calculated both analytically (Tijms 1994), and numerically (Trivedi 1982). Note that $p_{ijk}(0) = 1$ if $i = j = k$, and $p_{ijk}(0) = 0$ otherwise. Moreover, $\lim_{t\to\infty} p_{ijk}(t) = \pi_k$ if $j = 0$, and $\lim_{t\to\infty} p_{ijk}(t) = 0$ otherwise. Here, $\boldsymbol{\pi} = (\pi_0, ..., \pi_m)$ with $\Sigma_i \pi_i = 1$ denotes the vector of steady state probabilities.

In order to arrive at an expression for c_f and $c_p(\Delta t)$, we have to account for all initial states i, all minimal states $j \leq i$, and all final states $k \geq j$, with corresponding probabilities. Since $X(\cdot)$ is assumed to be restored into a stationary state after each preventive and corrective maintenance action, this yields the following expressions for c_f and $c_p(\Delta t)$:

$$c_f = \sum_{i=0}^{m} \pi_i \cdot k_f(i)$$

$$c_p(\Delta t) = \sum_{i=0}^{m} \sum_{j=0}^{i} \sum_{k=j}^{m} \pi_i \cdot k_p(j) \cdot p_{ijk}(\Delta t)$$

It is immediately clear that $c_p(\Delta t)$ is an analytic function, which decreases with

Δt. Moreover, $c_p(0) = \sum_i \pi_i \cdot k_p(i) < \sum_i \pi_i \cdot k_f(i) = c_f$, and $\lim_{\Delta t \to \infty} c_p(\Delta t) = k_p(0)$. In the simplest case, $c_p(\Delta t)$ is a linear combination of exponential functions in Δt.

Validation of the first stage model

As a starting point, let us determine the probability $P(X_t = Y_{\Delta t})$ that the minimal state during the interval $[0, \Delta t]$ is - amongst others - attained at time t, provided that $X(\cdot)$ is in a stationary state at time 0. Then, similar to the previous section, we have to account for all initial states i, all minimal states $j \leq i$, and all final states $k \geq j$, with corresponding probabilities. This yields the following expression for $P(X_t = Y_{\Delta t})$:

$$P(X_t = Y_{\Delta t}) = \sum_{i=0}^{m} \sum_{j=0}^{i} \sum_{k=j}^{m} \pi_i \cdot p_{ijj}(t) \cdot p_{jjk}(\Delta t - t)$$

It can easily be derived that for $m = 1$, this leads to an expression which is independent of t. Unfortunately, this property cannot be translated into a uniform distribution of the optimal starting time $\hat{t}$ over the maintenance interval $[0, \Delta t]$, since the final choice among the set of candidate starting times $\{0 \leq t \leq \Delta t \mid X_t = Y_{\Delta t}\}$ also depends on the selection strategy of the maintenance planner. But even if the maintenance planner chooses randomly among these candidate starting times, $\hat{t}$ would still not be uniformly distributed over $[0, \Delta t]$, see Appendix C for details. On the other hand, if we restrict ourselves to reversible Markov processes, and thus $\{X_t\}$ is stochastically identical to $\{X_{-t}\}$, it is immediately clear that the expected starting time for preventive maintenance equals $E\{\hat{t}\} = \Delta t/2$. Moreover, $\hat{t}$ is symmetrically distributed around this value.

A simulation study

In order to verify our first stage modelling assumptions, we carried out a simulation study for a queue-like production system with $m = 1$, $k_f(1) = 4$, $k_f(0) = k_p(1) = 2$, and $k_p(0) = 1$. After some elementary algebra, this yields the following expressions for $c_f = 2 \cdot \{1 + \frac{\lambda}{\lambda+\mu}\}$ and $c_p(\Delta t) = 1 + \frac{\lambda}{\lambda+\mu} \cdot e^{-\mu \cdot \Delta t}$, where we denote $\lambda = q_{01}$ and $\mu = q_{10}$ for notational convenience. In each simulation experiment, the optimal starting time $\hat{t} \in [t, t + \Delta t]$ for preventive maintenance was determined by choosing arbitrarily among the set of candidate starting times. In addition, the operating state $X(.)$ of the production system was not restored to its stationary state after each preventive and/or corrective maintenance action. The results are depicted as simulation (I) in Table 5.2.

Table 5.2: Optimal two-stage maintenance policies $\{t^*, \Delta t^*\}$ with corresponding costs $h(t^*, \Delta t^*)$ for a queue-like production system with $m = 1$, $k_f(1) = 4$, $k_f(0) = k_p(1) = 2$ and $k_p(0) = 1$, in comparison with 95% confidence intervals for the actual values based on simulation.

λ	μ	t^*	Δt^*	$h(t^*, \Delta t^*)$	simulation I	simulation II
1	1	0.568	0.373	2.257	2.251 ± 0.005	2.261 ± 0.005
1	5	0.609	0.278	1.725	1.764 ± 0.004	1.730 ± 0.005
1	10	0.661	0.191	1.642	1.647 ± 0.006	1.644 ± 0.003
5	1	0.500	0.474	2.672	2.428 ± 0.006	2.681 ± 0.007
5	5	0.499	0.386	1.936	1.933 ± 0.008	1.938 ± 0.007
5	10	0.574	0.266	1.767	1.773 ± 0.006	1.772 ± 0.006
10	1	0.488	0.491	2.764	2.552 ± 0.006	2.772 ± 0.006
10	5	0.464	0.414	2.024	2.014 ± 0.009	2.020 ± 0.005
10	10	0.538	0.288	1.836	1.835 ± 0.006	1.836 ± 0.006

From the simulation results, we conclude that $h(t^*, \Delta t^*)$ often comes close to the actual value based on simulation. On the other hand, there is a significant difference in each of the following cases: (i) $\lambda = 1$ and $\mu = 5$, (ii) $\lambda = 5$ and $\mu = 1$, and (iii) $\lambda = 10$ and $\mu = 1$. A closer look at the simulation results indicated that most problems were caused by the assumption that $X(.)$ is restored to a stationary state after each maintenance action. To this end, we carried out another simulation study (II) in which $X(.)$ was restored to a stationary state after each preventive and corrective maintenance action (see Table 5.2). Apparently, our two-stage maintenance policy must be handled with care if this stationarity assumption becomes too criticial. Simply stated, this means that the average sojourn times $\left|q_{ii}^{-1}\right|$ in each of the possible states $i \in \{0, ..., m\}$ must be relatively small in comparison with the average life time μ of the production system.

5.5 Computational results

To investigate the potential benefits of a two-stage generalized age maintenance policy, in relation to a classical age maintenance policy, we carried out a series of numerical experiments for a production system with Gamma distributed life times, with mean $\mu = 1$ and standard deviation $\sigma \in \{0.25, 0.50\}$. In each of these experiments, the maintenance cost functions were determined by $c_f = 100$ and $c_p(\Delta t) = \alpha + \beta \cdot e^{-\gamma \cdot \Delta t}$, where $\alpha \in \{10, 25\}$, $\beta \in \{5, 10\}$ and $\gamma \in \{1, 2, 5\}$. For each of the 24

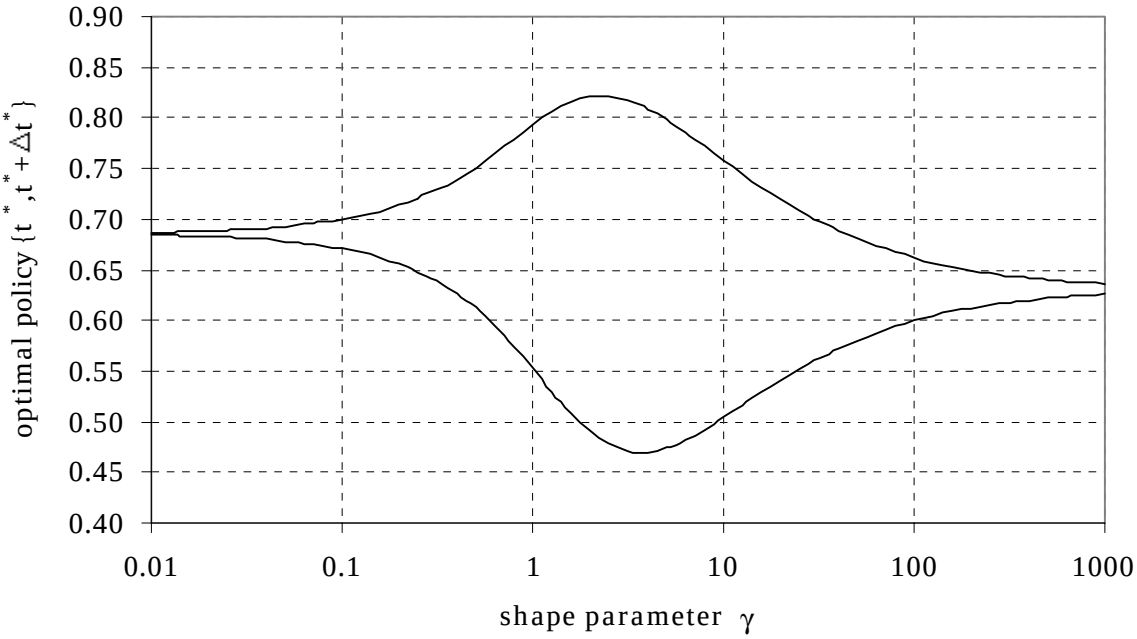


Figure 5.5: Optimal two-stage maintenance policies $\{t^*, t^* + \Delta t^*\}$ for a production system with $c_f = 100$, $c_p(\Delta t) = 25 + 10 \cdot e^{-\gamma \cdot \Delta t}$, and Gamma distributed life times ($\mu = 1$, $\sigma = \frac{1}{4}$), as a function of the shape parameter γ ($\Delta t_{\max} = \infty$).

test problems obtained this way, we determined the optimal age maintenance policy $t_0 < \infty$ with corresponding costs $g(t_0)$, and the optimal two-stage maintenance policy $\{t^*, \Delta t^*\}$ with corresponding costs $h(t^*, \Delta t^*)$. The results are depicted in Table 5.3.

As a starting point, we can see from Table 5.3 how the ratio of $h(t^*, \Delta t^*)$ and $g(t_0)$ varies with α, β and γ. In accordance with our general expectations, we observe that the relative improvement of a two-stage generalized age maintenance policy, compared with a classical age maintenance policy, increases with β and γ, but decreases with α. In a similar way, we can observe that the costs $h(t^*, \Delta t^*)$ of the optimal two-stage maintenance policy $\{t^*, \Delta t^*\}$ increases with α, β and σ, but decreases with γ.

Let us now take a closer look at the case $\sigma = 0.25$, $\alpha = 25$ and $\beta = 10$, and further investigate the behavior of t^* and Δt^* in relation to γ. If $\gamma \to 0$, the preventive maintenance cost function converges to $c_p(\Delta t) = \alpha + \beta = 35$ for all $\Delta t \geq 0$. In a similar way, $\gamma \to \infty$ leads to $c_p(\Delta t) = \alpha = 25$ for $\Delta t > 0$, whereas $c_p(0) = \alpha + \beta = 35$. In both cases, however, it follows from Theorem 2 that $\Delta t^* \to 0$. In other words, as γ tends to zero or infinity, the optimal two-stage maintenance policy coincides with the optimal age maintenance policy, i.e. $t^* \to t_0$ and $\Delta t^* \to 0$. For intermediate values of

Table 5.3: Optimal maintenance policies for a production system with Gamma distributed life times with mean $\mu = 1$ and standard deviation $\sigma \in \{0.25, 0.50\}$, corrective maintenance costs $c_f = 100$, and preventive maintenance costs $c_p(\Delta t) = \alpha + \beta \cdot e^{-\gamma \cdot \Delta t}$, for different values of σ, α, β and γ ($\Delta t_{max} = \infty$).

σ	α	β	γ	t_0	$g(t_0)$	t^*	Δt^*	$t^* + \Delta t^*$	$h(t^*, \Delta t^*)$	savings
0.50	25	10	1	0.721	86.4	0.440	0.525	0.965	83.7	3.2 %
0.50	25	10	2	0.721	86.4	0.390	0.551	0.941	80.9	6.4 %
0.50	25	10	5	0.721	86.4	0.393	0.434	0.827	77.2	10.7 %
0.50	25	5	1	0.636	80.6	0.485	0.293	0.778	79.7	1.2 %
0.50	25	5	2	0.636	80.6	0.436	0.368	0.804	78.4	2.8 %
0.50	25	5	5	0.636	80.6	0.424	0.339	0.763	76.2	5.5 %
0.50	10	10	1	0.489	65.3	0.264	0.410	0.674	61.5	5.8 %
0.50	10	10	2	0.489	65.3	0.202	0.472	0.674	56.9	12.9 %
0.50	10	10	5	0.489	65.3	0.192	0.399	0.591	50.0	23.4 %
0.50	10	5	1	0.421	55.2	0.300	0.229	0.529	54.0	2.3 %
0.50	10	5	2	0.421	55.2	0.247	0.311	0.558	51.9	6.1 %
0.50	10	5	5	0.421	55.2	0.225	0.308	0.532	48.0	13.2 %
0.25	25	10	1	0.686	60.3	0.553	0.240	0.793	58.8	2.6 %
0.25	25	10	2	0.686	60.3	0.491	0.331	0.822	56.2	6.9 %
0.25	25	10	5	0.686	60.3	0.474	0.325	0.799	51.5	14.7 %
0.25	25	5	1	0.658	53.4	0.588	0.132	0.720	53.0	0.9 %
0.25	25	5	2	0.658	53.4	0.544	0.206	0.750	52.0	2.7 %
0.25	25	5	5	0.658	53.4	0.516	0.240	0.756	49.7	7.0 %
0.25	10	10	1	0.603	38.3	0.432	0.289	0.721	36.1	5.8 %
0.25	10	10	2	0.603	38.3	0.337	0.407	0.744	32.5	15.2 %
0.25	10	10	5	0.603	38.3	0.327	0.382	0.709	26.7	30.3 %
0.25	10	5	1	0.572	30.0	0.477	0.170	0.647	29.3	2.4 %
0.25	10	5	2	0.572	30.0	0.414	0.262	0.676	27.9	7.0 %
0.25	10	5	5	0.572	30.0	0.383	0.289	0.672	25.0	16.8 %

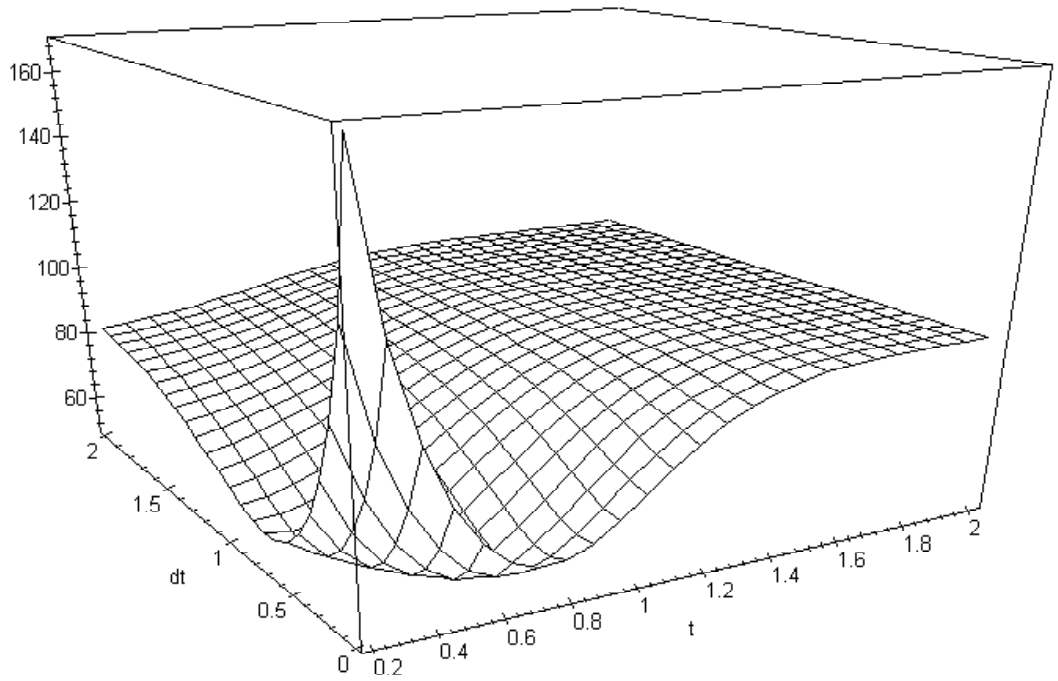


Figure 5.6: Two-stage maintenance policies $h(t, \Delta t)$ for a production system with Gamma distributed life times ($\mu = 1$, $\sigma = \frac{1}{4}$), $c_f = 100$ and $c_p(\Delta t) = 25 + 10 \cdot e^{-5 \cdot \Delta t}$: numerical optimization yields $t_0 \approx 0.686$ and $g(t_0) \approx 60.3$, whereas $t^* \approx 0.474$, $\Delta t^* \approx 0.325$ and $h(t^*, \Delta t^*) \approx 51.5$.

γ, different results are usually found (see Figure 5.5). An illustrative 3-dimensional plot of $h(t, \Delta t)$ in case $\gamma = 5$ is presented in Figure 5.6. The optimal age maintenance policy $t_0 \approx 0.686$ with costs $g(t_0) \approx 60.3$ is clearly visible along the t-axis.

5.6 Concluding remarks

In this chapter, we presented a two-stage generalized age maintenance policy, which more explicitly takes into account that (i) the initiation of preventive maintenance should be based on the technical and operating state of a production system, and (ii) this operating state is often known in advance over a finite rolling horizon in an operational planning phase. We presented a variety of modelling and optimization techniques, with which an optimal two-stage maintenance policy can be determined to a sufficient, and user-defined level of detail. Based upon a series of numerical

experiments, we concluded that significant savings can be obtained in comparison with a classical age maintenance policy.

Presumably, one of the main problems in real-life situations will be to express the expected preventive maintenance costs $c_p(\Delta t)$ as a function of the interval size Δt. Similarly, it might be difficult to come up with the actual distribution of $\hat{t}$ over the maintenance interval $[t, t + \Delta t]$. On the other hand, since the majority of practical applications in the maintenance area is based on ambiguous data (e.g. expert opinions, subjective data, fitted distributions), reasonable approximations should not cause major problems. From a practical point of view, optimal maintenance policies should always be handled with care; the two-stage maintenance policy is no exception to this rule.

Of course, our two-stage maintenance concept could also be applied to other maintenance models as well. For example, consider a classical **block replacement** model (Barlow and Proschan 1965), in which preventive maintenance is carried out as soon as t time units have elapsed since the last preventive maintenance action. In other words, preventive maintenance is carried out at times $\{t, 2 \cdot t, 3 \cdot t, ...\}$. Within our two-stage maintenance concept, this would mean that preventive maintenance is carried out somewhere between times t and Δt, somewhere between times $2 \cdot t$ and $2 \cdot t + \Delta t$, somewhere between times $3 \cdot t$ and $3 \cdot t + \Delta t$, etcetera. In line with this, the time between two consecutive preventive maintenance actions becomes a random variable, whose probability density function $\varphi_{t,\Delta t}(\cdot)$ is determined as follows:

$$\varphi_{t,\Delta t}(u) = \frac{1}{2 \cdot \Delta t} \cdot \left\{ 1 - \frac{|u - t|}{2 \cdot \Delta t} \right\} \quad , \quad t - 2 \cdot \Delta t \leq u \leq t + 2 \cdot \Delta t$$

Note that $\int_{t-2\cdot\Delta t}^{t+2\cdot\Delta t} \varphi_{t,\Delta t}(u)\ du = \frac{1}{2} \cdot 4 \cdot \Delta t \cdot (2 \cdot \Delta t)^{-1} = 1$ for all $t \geq 0$, $\Delta t > 0$. Let us now denote with $M(t)$ the expected cumulative deterioration costs (due to failures, repairs, etc.), t time units since the last preventive maintenance action. By doing this, we can derive the following expression for the expected maintenance costs per unit of time $h(t, \Delta t)$ associated with a two-stage generalized block replacement policy $\{t, \Delta t\}$. Here, we denote $\phi(u) = \varphi_{0,1}(u)$ for notational convenience:

$$h(t, \Delta t) = \frac{c_p(\Delta t) + \int_{t-2\cdot\Delta t}^{t+2\cdot\Delta t} \phi(\frac{u-t}{\Delta t}) \cdot M(u)\ du}{t}$$

Under some weak conditions, i.e. if $M(t)$ is strictly increasing and convex in t, it is possible to formulate similar procedures to arrive at the optimal two-stage generalized block replacement policy $\{t^*, \Delta t^*\}$. This and other possible model extensions, however, are left for future research.

5.7 Appendix

A Proof of Lemma 6

(a) Since $f(u) \equiv m(u)/n(u)$ is assumed to be strictly positive and strictly increasing in u, it follows that $\int_a^b m(u)\,du = \int_a^b f(u)\cdot n(u)\,du > f(a)\cdot\int_a^b n(u)\,du$. Similarly, it can be shown that $\int_a^b m(u)\,du < f(b)\cdot\int_a^b n(u)\,du$. Obviously, dividing by $\int_a^b n(u)\,du$ yields the desired result.

(b) Since $b = 0$ is trivial, we have to show that $g(a,b) \equiv \int_a^{a+b} m(u)\,du / \int_a^{a+b} n(u)\,du$ is strictly increasing in $a \geq 0$ for all $b > 0$. As a starting point, observe that (a) implies that $f(a) < g(a,b) < f(a+b)$. Let us now derive an expression for $g(a+h,b)$, where $h \to 0$:

$$
\begin{aligned}
g(a+h,b) &= \frac{\int_{a+h}^{a+b+h} m(u)\,du}{\int_{a+h}^{a+b+h} n(u)\,du} \approx \frac{\int_a^{a+b} m(u)\,du + h\cdot m(a+b) - h\cdot m(a)}{\int_a^{a+b} n(u)\,du + h\cdot n(a+b) - h\cdot n(a)} \\
&= \frac{g(a,b)\cdot\int_a^{a+b} n(u)\,du + h\cdot f(a+b)\cdot n(a+b) - h\cdot f(a)\cdot n(a)}{\int_a^{a+b} n(u)\,du + h\cdot n(a+b) - h\cdot n(a)} \\
&= g(a,b) + h\cdot\frac{(f(a+b)-g(a,b))\cdot n(a+b) + (g(a,b)-f(a))\cdot n(a)}{\int_a^{a+b} n(x)\,dx + h\cdot n(a+b) - h\cdot n(a)}
\end{aligned}
$$

It is clear that this yields the following expression for $\partial g(a,b)/\partial a$:

$$
\lim_{h\to 0}\frac{g(a+h,b)-g(a,b)}{h} = \frac{(f(a+b)-g(a,b))\cdot n(a+b) + (g(a,b)-f(a))\cdot n(a)}{\int_a^{a+b} n(x)\,dx}
$$

Since $f(a+b) > g(a,b)$, $n(a+b) > 0$, $g(a,b) > f(a)$, and $n(a) > 0$, it follows that $\partial g(a,b)/\partial a > 0$ for all $a \geq 0$, and thus $g(a,b)$ is strictly increasing in $a \geq 0$ for all $b > 0$.

(c) The proof is similar to (b).

B Analytical expressions for $g(t)$ and $h(t, \Delta t)$

The underlying observation behind these explicit formulas for $g(t)$ and $h(t, \Delta t)$, is that the following relations can be derived in case of Gamma distributed life times, with cumulative distribution function $\Gamma_{\alpha,\beta}(.)$, mean $\mu = \alpha \cdot \beta$, and variance $\sigma^2 = \alpha \cdot \beta^2$:

$$\int_0^t \Gamma_{\alpha,\beta}(u)\ du = t \cdot \Gamma_{\alpha,\beta}(t) - \alpha \cdot \beta \cdot \Gamma_{\alpha+1,\beta}(t)$$

$$\int_0^t u \cdot \Gamma_{\alpha,\beta}(u)\ du = \frac{1}{2} \cdot t^2 \cdot \Gamma_{\alpha,\beta}(t) - (\alpha + \alpha^2) \cdot \beta^2 \cdot \Gamma_{\alpha+2,\beta}(t)$$

C Distribution of $\hat{t}$ over $[0, \Delta t]$

For simplicity, and for notational convenience as well, we consider the case $m = 1$ with transition rates $q_{01} = q_{10} = q > 0$. As a starting point, let us denote with N_t the number of jumps of the process $X(\cdot)$ during the interval $[0, t]$, and elaborate upon the conditional probabilities $P(\hat{t} \leq t \mid N_{\Delta t} = n)$, with $n \in \mathbb{N}$. Since the maintenance planner chooses randomly among the set of candidate starting times $\{0 \leq t \leq \Delta t \mid X_t = Y_{\Delta t}\}$, it is immediately clear that $\hat{t}$ is uniformly distributed over $[0, \Delta t]$, given that no jumps occur during this interval:

$$P(\hat{t} \leq t \mid N_{\Delta t} = 0) = \frac{t}{\Delta t}$$

On the other hand, we have to account for all feasible realizations of $X(\cdot)$, in order to arrive at the averages relevant for $P(\hat{t} \leq t \mid N_{\Delta t} = 1)$. To this end, let us denote with $p_i(\tau)$ the probability density that $X_0 = i$ and the only jump during $[0, \Delta t]$ occurs at time τ. Then obviously, $P(\hat{t} \leq t, N_{\Delta t} = 1)$ can be expressed as follows:

$$P(\hat{t} \leq t, N_{\Delta t} = 1) = \int_0^t p_0(\tau) \cdot 1\ d\tau + \int_t^{\Delta t} p_0(\tau) \cdot \frac{t}{\tau}\ d\tau + \int_0^t p_1(\tau) \cdot \frac{t - \tau}{\Delta t - \tau}\ d\tau$$

Within this setting, it is easily derived that $p_0(\tau) = p_1(\tau) = \frac{1}{2} \cdot q \cdot e^{-q \cdot \Delta t}$ for all $\tau \in [0, \Delta t]$. Now substitution of these expressions in $P(\hat{t} \leq t, N_{\Delta t} = 1)$, and dividing by $P(N_{\Delta t} = 1)$, yields a non-uniform distribution of $\hat{t}$ over $[0, \Delta t]$, provided that exactly one jump occurs during this interval:

$$P(\hat{t} \leq t \mid N_{\Delta t} = 1) = \frac{t + \frac{1}{2} \cdot t \cdot \ln\left\{\frac{\Delta t}{t}\right\} + \frac{1}{2} \cdot (\Delta t - t) \cdot \ln\left\{\frac{\Delta t - t}{\Delta t}\right\}}{\Delta t}$$

Following a similar argument, it can be shown that $\hat{t}$ is uniformly distributed over $[0, \Delta t]$, provided that exactly two jumps occur during this interval. Based on explicit relations for $n = 1...7$, which were determined with the use of Maple, the following limiting behavior was observed:

$$P(\hat{t} \leq t \mid N_{\Delta t} = n) = \left\{ \begin{array}{ll} \frac{t}{\Delta t} & n = 0, 2, 4, 6, ... \\ \frac{t}{\Delta t} \cdot \frac{1+\ln \Delta t - \ln t}{2} + O(t^2 \log t) & n = 1 \\ \frac{t}{\Delta t} \cdot \frac{n}{n-1} + O(t^2 \log t) & n = 3, 5, 7, ... \end{array} \right\}$$

Since $P(N_{\Delta t} = n) > 0$ for all $n \in \mathbb{N}$ by definition, and $P(\hat{t} \leq t) = P(\hat{t} \geq \Delta t - t)$ for all $0 \leq t \leq \Delta t$, we conclude that $\hat{t}$ is not uniformly distributed over $[0, \Delta t]$, and even has singularities in $t = 0$ and $t = \Delta t$ of logarithmic type.

Chapter 6

Maintenance meets production at KLM Royal Dutch Airlines

In order to encounter some actual interactions between production and maintenance in a practical context, a case study has been carried out at the Line Maintenance department of KLM Royal Dutch Airlines. This department is responsible for the inspection, maintenance and repair of aircrafts during their stay at Schiphol Airport, as well as the assignment of aircrafts to flights within KLM's timetable. A decision support system has been developed with which maintenance managers are better equiped to determine how many maintenance slots of which type should be available in the timetable, and how many maintenance engineers of which type should be assigned to these slots, in order to satisfy the overall service levels set by higher management. The main objective of this study was to develop some fundamental and elementary queueing models, which could eventually assist maintenance managers in the formulation of several design criteria for KLM's timetables.

6.1 Introduction

KLM Royal Dutch Airlines has been the major Dutch airline since 1919. KLM's home base is Schiphol Airport nearby Amsterdam. Currently (1997), KLM owns about 90 aircrafts of 8 different types, which operate flights to and from about 150 cities in 80 countries. Traditionally, the safety of passengers and crew has had top priority in KLM's mission statement. Therefore, KLM carries out high-quality maintenance, relying on approximately 3000 employees in its overall maintenance department. To be specific, each aircraft is maintained preventively through major and minor inspections, and correctively during its stay at Schiphol Airport. Major inspections are

Table 6.1: Major inspection intervals for the intercontinental fleet.

inspection	# weeks	# flights	# flight hours
A check	6	150	650
C check	18	1.300	7.500
D check	72	5.200	25.000

performed in KLM's hangars after a certain amount of time, flights and/or flight hours, and their underlying structure is completely equivalent to the indirect clustering approach presented in chapter 3.

Minor inspections are conducted in between each arrival and departure at Schiphol Airport, and include a variety of so-called arrival, platform and departure services (Dijkstra et al. 1994). Amongst several other activities, arrival services consist of fixing ground power supply, compiling a list of technical complaints based on the crew's flight records, and collecting resources (e.g. mobile cranes and scaffoldings) for the platform services. Furthermore, platform services consist of checking the technical state of the aircraft, and performing repairs whenever necessary. Finally, departure services consist of a final technical check of the aircraft. In this chapter, we are mainly concerned with platform services, and **performing repairs** in particular.

These repair activities are carried out by employees of the Line Maintenance department. Currently, its workforce consists of approximately 250 highly-skilled and well-trained maintenance engineers. Their responsibility is to inspect, maintain, and repair KLM's aircrafts during their stay at Schiphol Airport. Due to internal and external safety rules, each maintenance engineer is licensed to carry out inspections on a limited number of aircraft types, and also has a specific skill for avionic resp. mechanical systems. The engineers obtain their licenses and skills by attending training programs consisting of theoretical and practical courses. Depending on their experience, it takes several months to several years to complete such a training program.

In general, the time required for arrival and departure services can be treated as a given constant, depending on the aircraft type. For similar reasons, the time required for preventive maintenance activities (platform check) is more or less fixed. The remaining period of ground time can be used for planned and/or unplanned corrective maintenance activities (i.e. performing repairs). It is determined by the difference between arrival and departure time of the aircraft under consideration, minus the required time for arrival services, departure services and platform checks. As such, this length can be derived from the underlying timetable operated by KLM Royal Dutch Airlines, in combination with the assignment of aircrafts to flights within

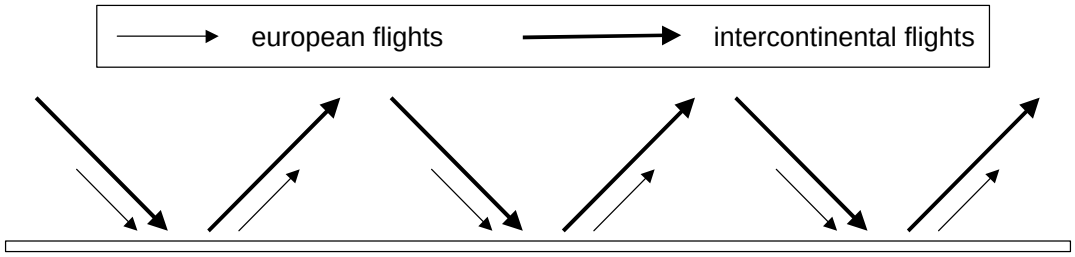


Figure 6.1: General structure of KLM's timetables.

this timetable. The latter decisions are also the responsibility of the Line Maintenance department.

Simply stated, KLM's **timetable** consists of a comprehensive collection of so-called city-city pairs with scheduled departure and arrival times, and corresponding aircraft type. Typically, this timetable follows a cyclical pattern, with a cycle length of exactly one week. On an average day, some clearly distinguishable peaks can be observed, caused by KLM's marketing strategy to minimize waiting times for transit passengers (connecting flights). Several times a day, a batch of intercontinental flights arrives at Schiphol Airport followed by a batch of continental flights to several destinations allover Europe, and vice versa (see Figure 6.1). In this study, we focus on the intercontinental fleet, which is mainly operated by the following aircraft types: Boeing 747-300 Combi (B743C), Boeing 747-300 Full Pax (B743P), Boeing 747-400 Combi (B744C), and Boeing 747-400 Full Pax (B744P). For a variety of technical, economical and operational reasons, these aircraft types are not mutually interchangeable, i.e. each aircraft type operates its own timetable.

The times between arrival and departure of KLM's aircrafts at Schiphol Airport are called **ground times** or **slots**. During these slots, several activities have to take place inside and outside the aircraft (e.g. cleaning, fueling, catering, boarding, etcetera), which usually take up to two hours on average. In the meanwhile, a post- and pre-flight inspection is conducted by specialized maintenance engineers. The remaining ground time is available for the elimination of (deferred) defects i.e. performing repairs. As a consequence, the minimal turn around time (i.e. if no corrective maintenance activities are carried out) equals approximately two hours, at least for the intercontinental fleet. On the other hand, the maximum allowed time for an aircraft to stay at its gate is restricted by Schiphol Airport, and equals approximately four hours. In that case, the aircraft must be transported to and from a buffer, each of which takes another half an hour on average. In view of efficiency, it is therefore important to incorporate as few as possible ground times between say 4 and 6 hours

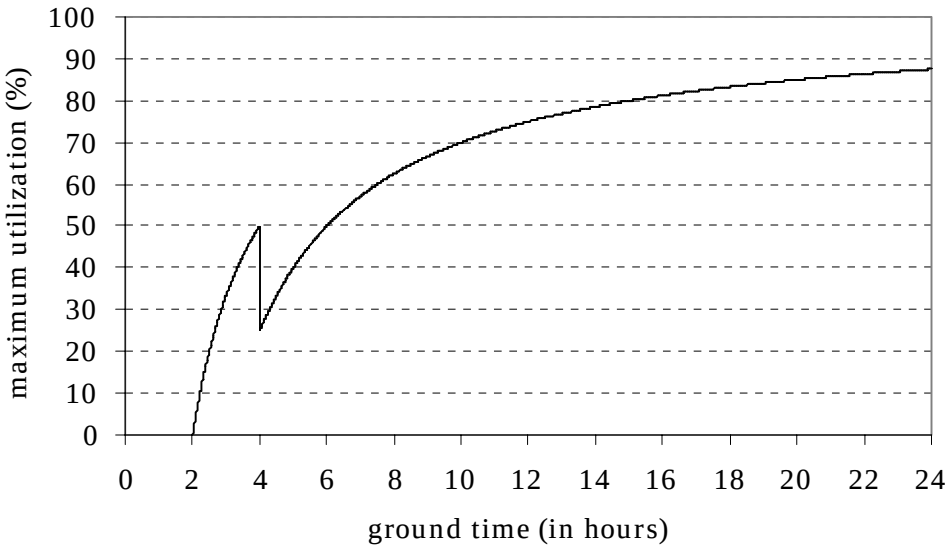


Figure 6.2: Maximum utilization for different ground times (slots) at Schiphol Airport, in terms of the fraction of available time that can be used for repair activities.

in the design of a timetable (see Figure 6.2). It was one of the main objectives of this study to provide maintenance managers with decision support in this respect.

Of course, the possibilities to keep the technical state of the aircrafts within the constraints set by higher management, are strongly related to the time that is reserved for the elimination of (deferred) defects. In this respect, it is not only the total amount of ground time (quantity), but also the relative frequencies of different slot types (quality) that counts. This effect is even stronger if one realizes that different defects may require different repair times (e.g. 3, 6 or 12 hours) and capacity (e.g. 1, 2 or 3 maintenance engineers), and may carry different due dates (e.g. 3, 10 or 30 days) as well. Another complicating factor in this respect is that each defect refers to a so-called **maintenance log** (aircraft vs. cabine) and **maintenance skill** (avionics vs. mechanics), which must be treated separately. In addition, technical no-go's i.e. defects that must be repaired before departure (zero due date) deserve special attention.

From a maintenance point of view, the timetable should provide enough opportunities to eliminate each defect before its due date, and within the amount of time and capacity that is available. On the other hand, such a timetable could never be optimal from an overall KLM perspective, since this would strongly reduce the number of scheduled flights, which is of particular commercial interest. Hence, the possibil-

ity of deferred defects, technical delays and cancellations is essential and inevitable in the design of KLM's timetables. In line with this, the performance of the Line Maintenance department is expressed in, and continuously monitored by two main elements. These are the technical dispatch dispunctuality (TDD), and the deferred defect list (DDL). In the remainder of this chapter, we will present some fundamental and elementary queueing models, which could provide maintenance managers with reasonable predictions and/or indicators of these performance measures at a strategical and tactical planning level.

The **technical dispatch dispunctuality** corresponds to the percentage of aircrafts that does not leave on time due to technical problems. In general, technical delays are due to deferred defects which cannot be repaired in time, and technical no-go's in particular. Depending on the aircraft type, a small amount of delay is usually allowed (15 minutes for the intercontinental fleet, 5 minutes for the continental fleet). Currently, the service level for KLM's technical dispatch dispunctuality is determined at a maximum of 4% for Schiphol Airport, and 2% world-wide. The **deferred defect list** corresponds to the collection of all reported defects that have not been eliminated yet, e.g. due to lack of time, capacity, spare parts and/or information. In this respect, a clear distinction is made between defects that are reported in the aircraft maintenance log (AML), and defects that are reported in the cabine maintenance log (CML). At the time this study was conducted, the service level for KLM's deferred defect list was determined at a maximum of 4 AML and 3 CML deferred defects per aircraft on average.

In the past few years, the maintenance department has used a tool called **critical flight analysis**, in order to determine the feasibility of a timetable with respect to the technical dispatch dispunctuality. So far, this tool has performed reasonably, and there is no specific reason for dramatic changes. Therefore, we will mainly focus on the accumulated workload associated with deferred defects. This does not necessarily mean, however, that there is no mutual relationship between these performance measures. After all, an increase in the number of deferred defects usually goes together with an increase in due date violations. In a similar way, an increase in due date violations implies an increase in technical delays and/or cancellations, and as such affects the technical dispatch dispunctuality.

The outline of this chapter is as follows. In section 6.2, we present a more detailed description of the problem under consideration, and discuss some related issues as well. Subsequently, a short introduction into our newly developed decision support system will be given in section 6.3, and we discuss the role of the underlying queueing models in some more detail. In section 6.4, we present a so-called time-based mod-

elling framework, which is mainly concerned with finding a proper match between large defects and large maintenance slots. In a similar way, section 6.5 comprises a so-called capacity-based modelling framework, which can assist maintenance managers in finding a proper match between overall workload and workforce, both expressed in man hours. In section 6.6, the results of our study are summarized, and some interesting opportunities for related research problems are discussed.

6.2 Problem description

Since the introduction of the 3-wave system a few years ago, and the more recent introduction of the SCORE (Schiphol COnnection REdesign) system (Bootsma 1997), the ground times at Schiphol Airport have been under constant pressure. In the past few years, this has resulted in a different workload for the Line Maintenance department, which on its turn has resulted in an increase of technical dispatch dispunctuality and deferred defects. At the same time, the goals of KLM's maintenance department are to increase punctuality, to decrease ground times, to reduce the number of deferred defects, and to raise productivity.

Under this pressure, the managers main problem is to find a good match between workload and workforce, all against reasonable costs in terms of the associated time and/or capacity. The elements that play a significant, if not crucial role in this match are timetables, failure rates and capacity profiles (see Figure 6.3). In the following section, these factors will be addressed in more detail. Here, we only mention that the quality of such a match is mainly determined by the number of **deferred defects**, and the number of **due date violations**. Too many due date violations may lead to unacceptable problems in an operational planning phase, in terms of delays and/or cancellations, and must therefore be controlled at a strategical and/or tactical level. In a similar way, too much deferred defects may result in poor aircraft quality, which is in conflict with the overall company objective to provide high quality service to its customers.

Planning for a good match of workload and workforce is therefore important, and it involves both strategical, tactical and operational planning issues. At the strategical level, management has to formulate a variety of design criteria, in order to arrive at a concept timetable (draft) which can be realized against reasonable costs, and within the constraints set. At the tactical level, management has to identify and solve (potential) problem areas within the draft, followed by a rough cut capacity planning for the maintenance department, in terms of the required number and type of maintenance engineers. At the operational level, management has to decide which aircraft

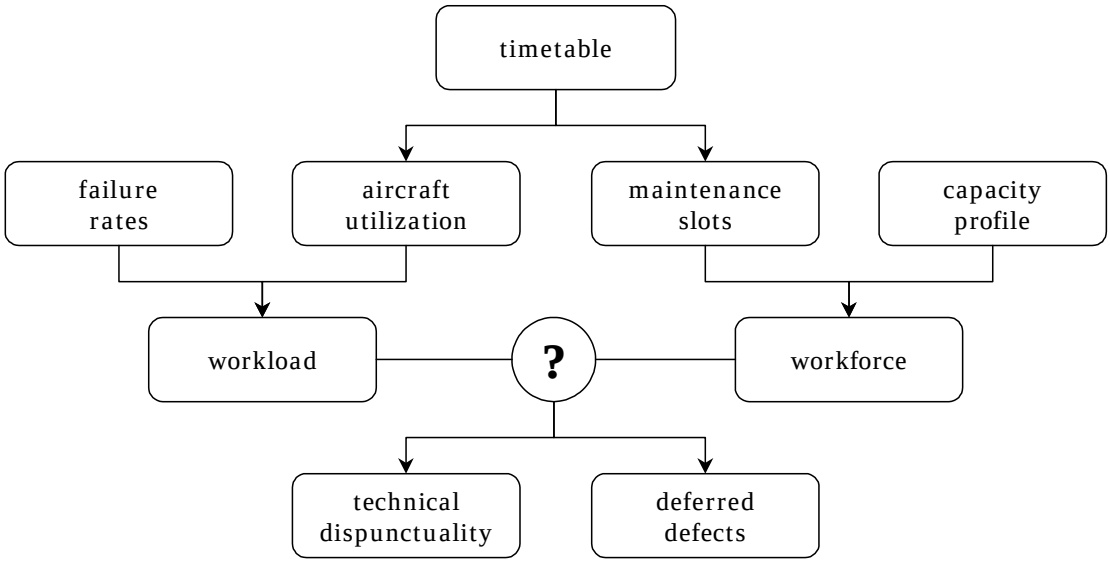


Figure 6.3: Crucial factors in the match between workload and workforce.

should operate which flight, and which maintenance activities should be carried out by which maintenance engineers during the resulting ground times. In this study, we focus at strategical and tactical planning issues, which means that the actual timing of maintenance slots in the timetable are not contained in our analysis. To be specific, the following decision problems are adressed:

- how many maintenance slots of which type must be available in the timetable?
- how many maintenance engineers of which type must be assigned to these slots?

Although the relative frequencies of different slot types are strongly related to the underlying structure of KLM's timetable, these figures can only be derived in an operational planning phase. After all, it depends on the day-to-day assignment of aircrafts to flight numbers, which slot types will actually be observed in practice, and with which relative frequencies (see Figure 6.4). At a strategical and tactical level, one often uses a last-in-first-out type of service discipline (LIFO) in order to derive the (expected) relative frequencies of different slot types within the timetable. By doing so, they usually overestimate the number of large slots, and underestimate the number of small slots.

The other way around, a first-in-first-out type of service discipline (FIFO) would normally underestimate the number of large slots, and overestimate the number of small slots. Throughout this study, we will assume that the relative frequencies of different slot types, as requested in the design of a timetable, can be realized in an

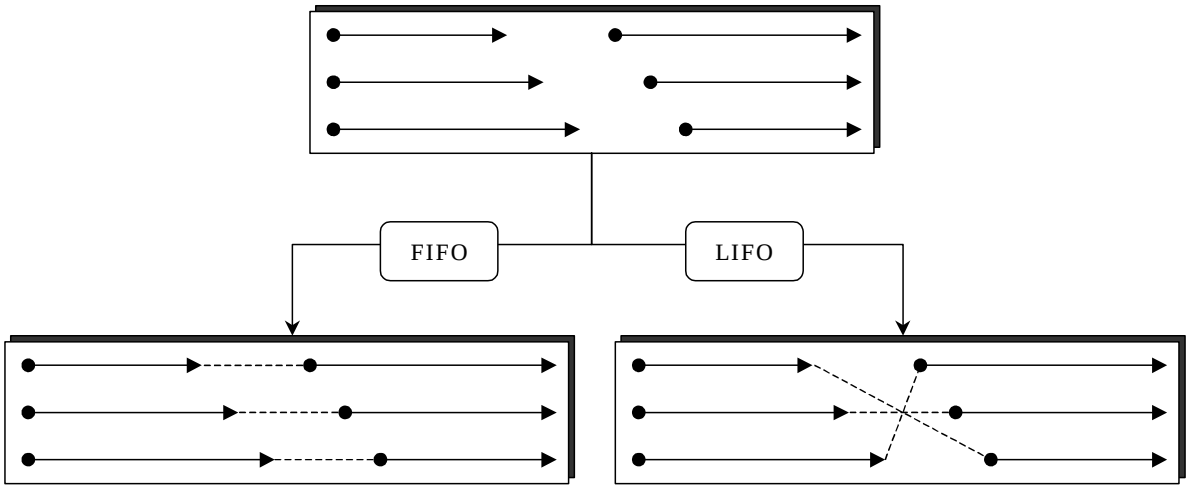


Figure 6.4: The relative frequencies of different slot types depend on the day-to-day assigment of aircrafts to flights within the timetable.

operational planning phase. In the newly developed SCORE system, the differences between both methods are in fact quite small, since most intercontinental flights arrive and depart in batches (see Figure 6.1). The reader is referred to Bootsma (1997) for a more detailed discussion on the structure of KLM's timetables.

Traditionally, major and minor inspections are planned in advance, and have never caused severe planning problems. On the contrary, defects are unplannable, and their stochastic nature goes together with fluctuations in the amount of workload offered to the maintenance department. Obviously, this phenomenon undermines the objective of eliminating each deferred defect before its due date, and within the required amount of ground time. Before we started this study, the managers based their design criteria with respect to the timetable merely on experience and intuition. Mathematical models were only based on long run averages, thereby neglecting the randomness of such events completely, e.g. see Owusu and Jessurun (1993) and Van der Eijck (1995). As a consequence, management could not evaluate or foresee potential problems associated with a (conceptual) timetable to a sufficient level of detail.

In the past few years, and in accordance with the second performance criterion, management has mainly been focussed on reducing the number of deferred defects. Therefore, small defects have usually received top priority in the operational planning and scheduling of maintenance activities. Simply stated, this priority setting strategy is based on the well-known shortest-processing-time (SPT) first service discipline (Silver et al. 1998), which aims at minimizing the average waiting time per

Table 6.2: Average repair times of deferred defects: 99% confidence intervals expressed in minutes (Van der Eijck, 1995).

	B743C	B743P	B744C	B744P
Aircraft Maintenance Log	84 ± 10	78 ± 8	87 ± 9	83 ± 11
Cabine Maintenance Log	35 ± 3	37 ± 4	36 ± 3	38 ± 4

defect. In line with this idea, large slots have often been used for the elimination of multiple small defects, rather than a few larger ones. Obviously, this cannot be the right planning strategy, since opportunities for large defects are scarce, whereas opportunities for small defects are (relatively) numerous. Throughout this study, we decided to reformulate the second performance criterion in terms of the workload associated with deferred defects, rather than the number of deferred defects itself.

In our view, this alternative performance criterion would lead to an important change in KLM's attitude towards corrective maintenance planning, and would eventually increase the overall performance of KLM Royal Dutch Airlines. At least, it served as a starting point for our analysis. Based on the average processing times of deferred defects within the intercontinental fleet (see Table 6.2), we proposed the following target levels: approximately 5 AML and 2 CML deferred man hours per aircraft on average.

6.3 Decision support system

In view of the addressed problem areas, the management of KLM's maintenance department had the impression that the quality of decision making could be improved by the introduction of a decision support system. Simply stated, the decision support system that we developed consists of four main elements, viz. a timetable module, a workload module, a workforce module, and an analysis module. With this decision support system, maintenance managers are better equiped to determine how many maintenance slots of which type should be available in the timetable, and how much capacity of which type should be assigned to these slots, in order to fulfil the overall company objectives within the constraints set by higher management. In the remainder of this section, a brief description of each module will be given.

As a starting point, the **timetable module** contains several fundamental characteristics of KLM's timetable for a specific aircraft type. These are (i) the number of aircrafts, (ii) the average number of flights per aircraft per day, (iii) the average number of flight hours per flight, (iv) the average number of flights per cycle, and (v)

the relative frequencies of different slot types. Here, a cycle is defined as a subsequent departure from and arrival at Schiphol Airport. This is an important quantity, since the majority of defects - the larges ones in particular - can only be tackled at Schiphol Airport, and not at so-called outbound stations. In general, most cycles consist of two flights, i.e. to and from a specific destination. Nevertheless, cycles of three or even more flights do also occur in practice, especially in the intercontinental fleet.

Secondly, the **workload module** contains the arrival rates of defects per flight and per flight hour, thereby making a clear distinction between different categories in terms of (i) the underlying maintenance log, (ii) the corresponding maintenance skill, (iii) the required repair time and capacity, and (iv) the associated due date. The underlying observation behind these so-called failure rates is that some deterioration processes are flight dependent (e.g. motors, engines), some are flight-hour dependent (e.g. chairs, lights), and some are a combination of both. Nevertheless, a preliminary study by Van der Eijck (1995) pointed out that the majority of failures, at least in terms of the associated workload, is flight-dependent. Eventually, a **database module** should be incorporated, which keeps track of all flights, flight hours, and defects in the intercontinental fleet. By doing this, it might become possible to determine the above-mentioned failure rates automatically, e.g. by using a multiple regression technique (Dunn and Clark 1974).

Finally, the **workforce module** determines how much maintenance engineers of which type (avionics and/or mechanics) should be assigned to each type of slot in the timetable. By doing this, the user can investigate the consequences of different capacity profiles at a strategical planning level (e.g. high capacity on large slots and low capacity on small slots, or vice versa). From now on, a complete set of figures concerning data on the timetable, data on the workload, and data on the workforce, is called a scenario. The **analysis module** provides the user with extensive possibilities for analyzing different scenarios. It consists of routines which estimate the workload and workforce per week, and evaluate the quality of the match between workload and workforce. In the following sections, these routines will be addressed in more detail, as well as the underlying queueing models and assumptions. Here, we only mention that a clear distinction has been made with respect to time and capacity.

Simply stated, the timetable must provide enough time (large slots) in order to cope with large defects, and enough capacity (manhours) in order to cope with all defects. In line with this idea, we have developed a separate **time-based** and **capacity-based** modelling framework, in order to define the quality of the match between workload and workforce. The basic underlying motivation behind each modelling framework is that workload accumulates until the arrival of workforce, but

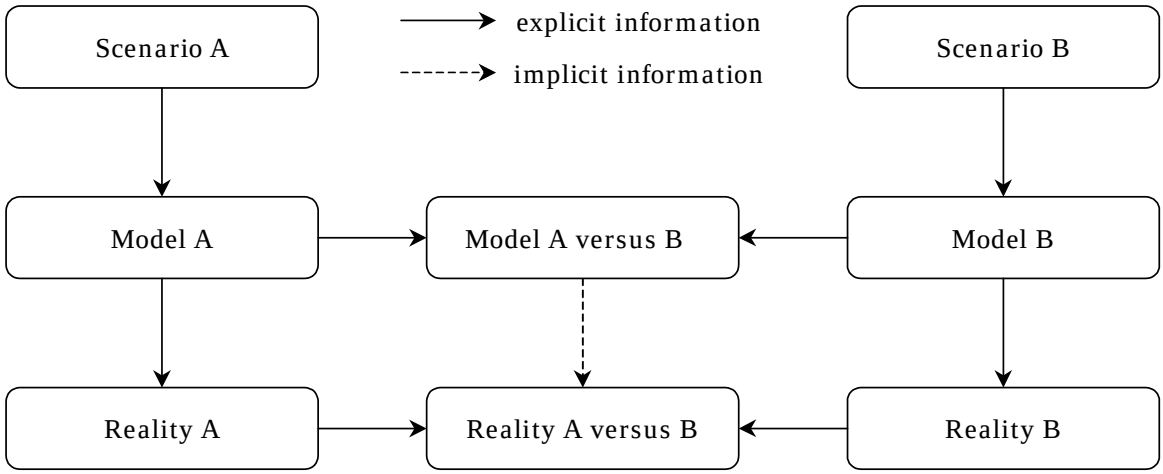


Figure 6.5: Scenario analysis via implicit and explicit information flows.

workforce cannot accumulate until the arrival of workload. Since these arrivals take place according to complex (stochastic) processes, the existence of deferred defects, and thus the possibility of due date violations, is inevitable. In line with this idea, management is primarily interested in the average amount of deferred workload, as well as the probability of due date violation for different types of defects. The analysis module provides reasonable estimates and/or indicators for these and other performance measures within a given scenario.

With this decision support system, it is of course also possible to compare different scenarios with each other in a strategical planning phase. In fact, and as long as our models have not been verified with actual data, this is exactly what our decision support system should be used, and was designed for. More specifically, our model outcomes for a specific scenario should be handled with care if interpreted explicitly, but may provide useful implicit information in relation to other model outcomes for other scenarios (see Figure 6.5).

From a practical point of view, this means that our models could be used to assist maintenance managers with comparative studies into alternative timetables. This is a potentially valuable insight, since there is still little on-the-job experience with our decision support system, and validation and/or modification of our models is yet to come. Nevertheless, we believe that they contain some interesting features, which are worth mentioning here. In the following sections, we will briefly describe the underlying queueing models and assumptions for our time-based and capacity-based modelling framework.

6.4 Time-based modelling framework

As a starting point, the time-based modelling framework is concerned with finding a proper match between large defects on the one hand, and large maintenance slots on the other hand. Since large slots should primarily be used for large defects, and as such should be treated with the highest priority, we assume that there is always enough manpower available to eliminate large defects. Therefore, the main ingredients of our time-based modelling framework are the complex stochastic processes associated with the arrivals of large defects and large slots.

6.4.1 Input and output specifications

As a starting point, the arrival rates of large defects are derived from the data provided by the timetable and workload modules, thereby making a clear distinction between different categories in terms of the required maintenance slot and the corresponding due date. In accordance with current KLM practices, we used 4 categories for large slots (>24 hours, 16-24 hours, 12-16 hours and 6-12 hours) and 4 categories for due dates (<1 day, 1-3 days, 3-10 days, and 10-120 days). Recall that maintenance slots of less than 6 hours can hardly be used efficiently, especially for large defects (see Figure 6.2). Nevertheless, our modelling framework could easily be generalized to cope with more (or less) categories.

As a next step, the average number of large slots per week is calculated from the data specified in the timetable module, again for each of the above-mentioned categories. Based upon these figures, the decision support system provides the user with reasonable estimates of the average waiting times for different types of defects, as well as the corresponding probabilities of due date violation. Subsequently, it determines a reasonable estimate of the average number of due date violations per week, which obviously is a useful performance indicator in view of comparing different scenarios with each other.

6.4.2 Model and assumptions

As we explained before, due dates are categorized into m different types, and maintenance slots into n different types. With d_i and t_j, we denote the typical or average due date and slot size associated with type i resp. type j. For notational convenience, and without loss of generality, we assume that $d_1 < ... < d_m$ and $t_1 > ... > t_n$. To continue our analysis, we denote with λ_{ij} the arrival rate of type (i, j) defects, i.e. defects with a due date of type i, that require a slot of type j. In a similar way,

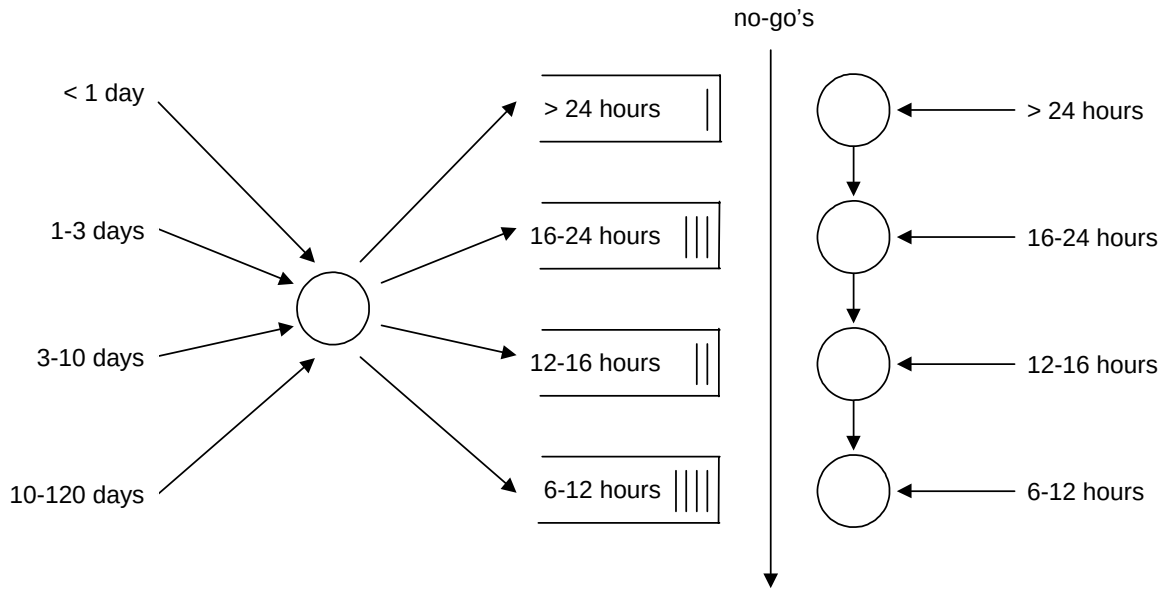


Figure 6.6: General structure of the time-based modelling framework

we let μ_j denote the arrival rate of type j slots. Finally, π_j reflects the probability that a slot of type j cannot be used for planned maintenance, e.g. due to a technical no-go or other causes. Currently, and in line with the above-mentioned modelling framework, KLM operates $m = 4$ different due date types, and $n = 4$ different slot types (see Figure 6.6).

Since there is no specific information available about the actual timing of large maintenance slots within the timetable, our modelling framework is based on the assumption that large defects and large slots arrive according to mutually independent Poisson processes. For similar reasons, we assume that each maintenance slot can be used for the elimination of at most one single defect. The question now remains which defect should be assigned to which maintenance slot. In general, this assignment should be based on the duration of the slot, and the due date of each defect. Since large defects should receive top priority, we adopted the following (approximate) priority rule at this strategical/tactical level: amongst all defects with the largest but still appropiate slot type, select the one with the lowest due date type. For example, a type 3 slot (12-16 hours) examines the collection of deferred defects in the following sequence: $(1,3) \rightarrow (2,3) \rightarrow (3,3) \rightarrow (4,3) \rightarrow (1,4) \rightarrow (2,4) \rightarrow (3,4) \rightarrow (4,4)$.

Our analysis now proceeds as follows. As a starting point, we denote with $\tilde{\lambda}_1 = \lambda_{11} + ... + \lambda_{m1}$ the arrival rate of defects that require a slot of type 1, and with $\tilde{\mu}_1 = \mu_1 \cdot \pi_1$ the arrival rate of such slots. By doing this, the waiting time of type $(i,1)$ defects is equivalent to the sojourn time in a single-server preemptive priority queue,

with exponentially distributed interarrival and service times for each priority class, in which the service of a lower priority job is interrupted as soon as a higher priority job enters the system. According to White and Christie (1958), this means that the following expression can be derived for the first two moments $E\{W_{i1}\}$ and $E\{W_{i1}^2\}$ of the waiting time W_{i1} for type $(i,1)$ defects. Here, we denote $\tilde{\rho}_{i1} = \lambda_{i1}/\tilde{\mu}_1 < 1$ and $\tilde{\sigma}_{i1} = \sum_{k \leq i} \tilde{\rho}_{k1} < 1$ for notational convenience:

$$E\{W_{i1}\} = \frac{1}{\tilde{\mu}_1} \cdot \frac{1}{(1-\tilde{\sigma}_{i-1,1}) \cdot (1-\tilde{\sigma}_{i1})}$$

$$E\{W_{i1}^2\} = \frac{2}{\tilde{\mu}_1^2} \cdot \left\{ \frac{1}{(1-\tilde{\sigma}_{i-1,1})^2 \cdot (1-\tilde{\sigma}_{i1})^2} + \frac{\tilde{\sigma}_{i-1,1}}{(1-\tilde{\sigma}_{i-1,1})^3 \cdot (1-\tilde{\sigma}_{i1})} \right\}$$

Let us now take a closer look at defects of type $(i,2)$, i.e. defects which require a slot of type 1 or 2. Since slots of type 1 are primarily used for type $(i,1)$ defects, it is easily verified that the arrival rate of these slots for type $(i,2)$ defects equals $\tilde{\mu}_2 = \tilde{\mu}_1 \cdot (1-\tilde{\rho}_1) + \mu_2 \cdot \pi_2$, where $\tilde{\rho}_1 = \tilde{\rho}_{11} + ... + \tilde{\rho}_{m1} = \tilde{\lambda}_1/\tilde{\mu}_1$. Clearly, the first term refers to slots of type 1 which are not used for type $(i,1)$ defects, whereas the second terms refers to slots of type 2. Our analysis now proceeds by assuming that the waiting time of type $(i,2)$ defects can also be modelled as the sojourn time in a single-server preemptive priority queue. Of course, this reasoning can only hold approximately, since the arrival process of type 1 slots for type 2 defects is no Poisson process in general. In an analagous way, we can now approximate the first two moments $E\{W_{i2}\}$ and $E\{W_{i2}^2\}$ of the waiting time W_{i2} for type $(i,2)$ defects. Again, we denote $\tilde{\rho}_{i2} = \lambda_{i2}/\tilde{\mu}_2 < 1$ and $\tilde{\sigma}_{i2} = \sum_{k \leq i} \tilde{\rho}_{k2} < 1$ for notational convenience:

$$E\{W_{i2}\} = \frac{1}{\tilde{\mu}_2} \cdot \frac{1}{(1-\tilde{\sigma}_{i-1,2}) \cdot (1-\tilde{\sigma}_{i2})}$$

$$E\{W_{i1}^2\} = \frac{2}{\tilde{\mu}_2^2} \cdot \left\{ \frac{1}{(1-\tilde{\sigma}_{i-1,2})^2 \cdot (1-\tilde{\sigma}_{i2})^2} + \frac{\tilde{\sigma}_{i-1,2}}{(1-\tilde{\sigma}_{i-1,2})^3 \cdot (1-\tilde{\sigma}_{i2})} \right\}$$

For $j > 2$, similar results can be obtained. Further details are skipped, since they are not so relevant for what follows. Our analysis is now based on the approximate reasoning that the waiting time W_{ij} of type (i,j) defects is a Gamma distributed random variable, with known parameters $\alpha_{ij} = E\{W_{ij}\}^2/Var\{W_{ij}\}$ and $\beta_{ij} = Var\{W_{ij}\}/E\{W_{ij}\}$, where $Var\{W_{ij}\} = E\{W_{i1}^2\} - E\{W_{i1}\}^2$. Summarizing, this leaves us with the following approximation for the probability of due date violation

Table 6.3: A comparitive study of 9 scenarios within the time-based modelling framework: average probability of due date violation per defect, in relation to the relative frequencies of different slot types (numerical example based on imaginary data).

slot type	I	II	III	IV	V	VI	VII	VIII	IX
>24 hours	2 %	2 %	2 %	2 %	4 %	4 %	4 %	4 %	6 %
16-24 hours	2 %	2 %	2 %	4 %	4 %	4 %	4 %	6 %	6 %
12-16 hours	4 %	4 %	6 %	6 %	6 %	6 %	8 %	8 %	8 %
6-12 hours	8 %	10 %	10 %	10 %	10 %	12 %	12 %	12 %	12 %
% too late	37.5 %	26.0 %	17.2 %	11.1 %	7.1 %	6.7 %	5.9 %	4.9 %	3.6 %

for type (i, j) defects. Here, $\Gamma_{\alpha,\beta}(.)$ denotes the cumulative distribution function of a Gamma distributed random variable with mean $\alpha \cdot \beta$ and variance $\alpha \cdot \beta^2$:

$$P\{W_{ij} > d_i\} = 1 - \Gamma_{\alpha_{ij},\beta_{ij}}(d_i)$$

Together with the arrival rates λ_{ij} of type (i, j) defects, this comes down to an arrival rate $\sum_{i,j} \lambda_{ij} \cdot P\{W_{ij} > d_i\}$ of due date violations. In our view, this value could be of considerable interest in comparing different scenarios at a strategical planning level, in view of the match between large slots and large defects. Eventually, the management of KLM's maintenance department could set a treshold value, based on historical data and expert opinions. Moreover, they could use this time-based modelling framework in the identification of potential problem areas within a timetable, by taking a closer look at $P\{W_{ij} > d_i\}$ and/or $\lambda_{ij} \cdot P\{W_{ij} > d_i\}$ for all i, j.

6.4.3 Numerical example

To illustrate our newly developed time-based modelling framework, we evaluated 9 different scenarios for the relative frequencies of different slot types in one of KLM's timetables. The results of this scenario analysis are depicted in Table 6.3. As we expected, the percentage of large defects which cannot be repaired before their due date, is strongly related to the relative frequencies of different slot types. For example, a closer look at scenarios I and VII indicates that the number of due date violations is reduced with a factor 6, if the number of large maintenance slots is increased with a factor 2. Obviously, such figures might provide maintenance managers with potentially valuable quantitative insights, which were previously not available.

6.5 Capacity-based modelling framework

Our capacity-based modelling framework is concerned with finding a proper match between workload and workforce in terms of manhours. In this respect, a clear distinction must be made between **released** and **unreleased** workload. Simply stated, released workload refers to deferred defects that are waiting to be scheduled, whereas unreleased workload consists of deferred defects that cannot be scheduled yet (see Figure 6.7). In general, the existence of unreleased workload is due to e.g. lack of information, equipment and/or materials, whereas released workload is mainly caused by lack of time and/or capacity. Obviously, the unreleased workload cannot be reduced by providing more time and/or capacity in the timetable, since these defects are due to external factors. This is a potentially valuable insight, which is further exploited in our model.

6.5.1 Input and output specifications

First of all, information must be available with respect to the average workload per cycle expressed in man hours, thereby making a clear distinction between the required maintenance skill (avionics/mechanics) and the corresponding maintenance log (aircraft/cabine). Subsequently, the user must provide the relative frequencies of different slot types in the timetable, as well as the capacity in terms of the number of avionic and/or mechanic maintenance engineers that are assigned to these slots. Finally, the user must specify how much time it takes (on average) for a defect to become part of the released workload, the so-called **external lead time**. In addition, the user can specify the minimal turn around time and maximal gate time, which determine the relation between ground time and available repair time (see Figure 6.2). Based upon these figures, the decision support system provides the user with an estimate of the average number of deferred man hours, for the aircraft maintenance log as well as the cabine maintenance log.

6.5.2 Model and assumptions

As a starting point of our analysis, and in accordance with our time-based modelling framework, we assume that the arrival of defects can be modelled as a homogeneous Poisson process. In order to determine the average amount of **unreleased workload**, it is now completely natural to assume that there is no correlation between the external lead times of different defects. By doing so, the unreleased workload can be modelled as a $M/G/\infty$ queue. Using Little's formula, it is easily verified that the

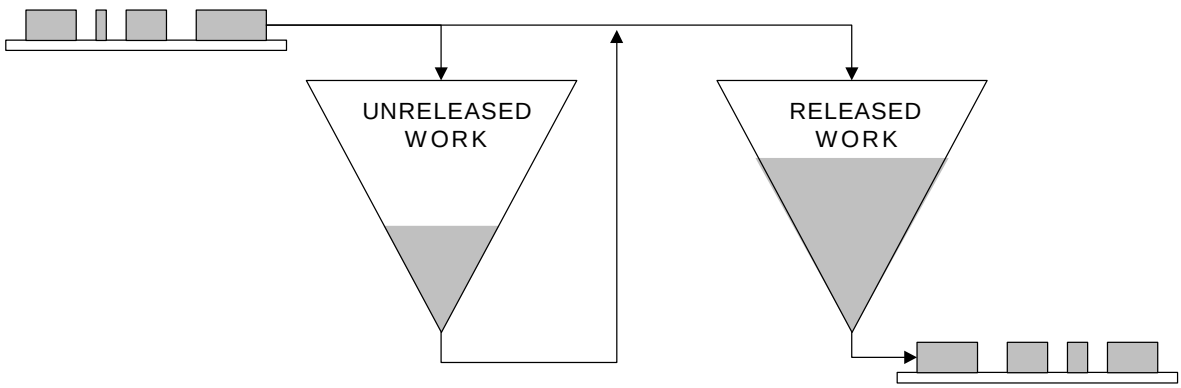


Figure 6.7: General structure of the capacity-based modelling framework

average unreleased workload equals λ/μ, where λ denotes the mean arrival rate of workload (e.g. 20 manhours per day), and $1/\mu$ denotes the average external lead time (e.g. 1 day).

Our model for **released workload** is now based on the assumption that there is no correlation between the workload and workforce per cycle. In other words, workload and workforce are modelled as mutually independent stochastic processes, one that degrades resp. one that upgrades the technical state of the aircraft. More specifically, the workload and workforce per cycle are modelled as independent stochastic variables Y and Z respectively, with known distribution functions. Within this setting, it is immediately clear that the deferred workload X just before departure must satisfy the following balance equation. Here, we define $[x]^+ = \max\{0, x\}$ for notational convenience:

$$X = [X + Y - Z]^+$$

This equation is known as Lindley's equation (Lindley 1952), for which explicit solutions are not readily available. Nevertheless, explicit solutions can be found for $M|G|1$ and $G|M|1$ queueing models, see e.g. Heyman and Sobel (1982). More specifically, $P(X \leq x)$ and thus $E\{X\}$ can be determined analytically, if Y is exponentially distributed with mean $1/\mu$:

$$P(X \leq x) = 1 - \alpha \cdot e^{-\mu \cdot x \cdot (1-\alpha)} \Rightarrow E\{X\} = \frac{\alpha}{\mu \cdot (1-\alpha)}$$

Here, $0 < \alpha < 1$ is the unique solution to $x = \mathcal{Z}(\mu \cdot (1-x))$, where $\mathcal{Z}(.)$ denotes the Laplace-Stieltjes transformation of $G(.)$, and $G(z) = P(Z \leq z)$ denotes the cumulative distribution function of the workforce per cycle:

$$\mathcal{Z}(s) = E\{e^{-s\cdot Z}\} = \int_0^\infty e^{-s\cdot z}\, dG(z) = \int_0^\infty g(z)\cdot e^{-s\cdot z}\, dz$$

The question remains how to determine $g(z)$. To this end, recall that the user must define a collection of different slot types, as well as the relative frequencies of occurence, and the capacities assigned to these slots. The decision support system converts these figures into a finite set of m different workforce classes $[a_i, b_i]$, expressed in manhours, with relative frequencies of occurence p_i. To achieve this, it uses the data with respect to minimal turn around time and maximal gate time (e.g. 2 resp. 4 hours in Figure 6.2). Subsequently, and in line with the above, we can define $g(z)$ rather straightforwardly as follows:

$$g(z) = \sum_{i:a_i\le z\le b_i} \frac{p_i}{b_i - a_i}$$

With this in mind, we can easily derive an analytical expression for $\mathcal{Z}(s)$. As a consequence, α can be determined numerically to a sufficient level of detail, using standard search techniques:

$$\mathcal{Z}(s) = \int_0^\infty g(z)\cdot e^{-s\cdot z}\, dz = \sum_{i=1}^m \int_{a_i}^{b_i} \frac{p_i}{b_i - a_i}\cdot e^{-s\cdot z}\, dz = \frac{1}{s}\cdot\sum_{i=1}^m p_i\cdot\frac{e^{-s\cdot a_i} - e^{-s\cdot b_i}}{b_i - a_i}$$

Of course, the outcomes of our models will strongly overestimate the actual values observed in practice. First of all, and in order to arrive at an analytical expression for $E\{X\}$, we modelled the workforce per cycle as an exponential distribution, whereas a Poisson or normal distribution would certainly be more realistic. Secondly, the workload per cycle Y and the workforce per cycle Z were modelled as mutually independent random variables, and we can probably do much better in practice. In our decision support system, this has been accounted for by multiplying our model outcomes with a **correction factor**, such that the predicted values would correspond with reality. Unfortunately, and due to lack of information and available time, we have not been able to come up with a properly validated estimation of this correction factor. So far, we have used a rough impression of this correction factor instead. Nevertheless, our model was still considered as a useful tool for comparing different scenarios for slot type distributions and capacity profiles.

Table 6.4: A comparitive study of 7 different scenarios within the capacity-based modelling framework: average slot length, ground time, and deferred workload, in relation to the relative frequencies of different slot types (numerical example based on imaginary data).

slot type	I	II	III	IV	V	VI	VII
24-32 hours	5 %	4 %	4 %	4 %	4 %	4 %	10 %
16-24 hours	5 %	4 %	4 %	4 %	4 %	10 %	4 %
12-16 hours	5 %	4 %	4 %	4 %	10 %	4 %	4 %
6-12 hours	10 %	9 %	9 %	15 %	9 %	9 %	9 %
4-6 hours	25 %	24 %	30 %	24 %	24 %	24 %	24 %
2-4 hours	50 %	55 %	49 %	49 %	49 %	49 %	49 %
slot length	6.75	6.14	6.26	6.50	6.80	7.16	7.64
ground time	4.25	3.69	3.75	3.99	4.29	4.65	5.13
unreleased AML	1.46	1.46	1.46	1.46	1.46	1.46	1.46
released AML	3.50	5.08	4.85	3.49	3.05	2.88	2.82
total AML	4.96	6.54	6.31	4.95	4.51	4.34	4.27
unreleased CML	0.68	0.68	0.68	0.68	0.68	0.68	0.68
released CML	1.63	2.37	2.26	1.63	1.42	1.34	1.31
total CML	2.31	3.05	2.94	2.31	2.10	2.03	2.00

6.5.3 Numerical example

To illustrate our newly developed capacity-based modelling framework, we evaluated 7 different scenarios for the relative frequencies of different slot types in one of KLM's timetables. In each scenario, the numbers of avionic resp. mechanic maintenance engineers assigned to each slot type were fixed. The results of this scenario analysis are depicted in Table 6.4. As we expected, an increase in the average length of maintenance slots usually goes together with a decrease in (released) deferred workload. On the other hand, a closer look at scenarios I and IV shows that it is also possible to improve the performance in this respect, while at the same time reducing the average length of maintenance slots. The underlying reasoning behind this counter-intuitive behavior is that - once again - frequent and short interruptions of the production process are to be preferred above infrequent and long ones, all other things being equal. Since maintenance slots of 2-6 hours can hardly be used efficiently (see Figure 6.2), this means that the relative frequency of 6-12 hour maintenance slots should be increased. This explains the attractiveness of scenario IV in relation to all other scenarios.

6.6 Concluding remarks

Of course, the decision support system that we developed is still far from providing absolute answers to relevant questions. After all, we made a lot of assumptions in order to arrive at explicit formulas for a variety of useful performance indicators, and verification and/or modification of these models and assumptions is yet to come. Nevertheless, we believe that our decision support system contains some interesting elements, with which maintenance managers are better equiped to determine how many slots of which type must be available in the timetable, and how many maintenance engineers must be assigned to these slots. In addition, our models could also be used to provide valuable support in each of the following dimensions:

- the impact of new time table structures,
- the effect of aircraft fleet derioration,
- the influences of due date adjustments,
- the benefits of external lead time reductions.

Summarizing, the decision support system provides the management of KLM's Line Maintenance department with information that was previously not available. It increases their insight into various strategical and tactical problems that must be solved within the maintenance department. On the other hand, we should keep in mind that the results obtained with our decision support system are based on approximate modelling techniques, and as such must be handled with care. Therefore, the user of our decision support system must judge the practical value of the model outcomes in light of considerations that were not explicitly accounted for.

To conclude this chapter, let us now briefly discuss the need for a decision support system at the operational planning level. Simply stated, the maintenance department must decide (i) which aircrafts should operate which flights, (ii) which capacity should be assigned to the resulting ground times, and (iii) which defects should be eliminated with this capacity. In general, these decisions relate to each other in a very complex manner. Moreover, they are restricted by several additional constraints, either economical, technological, combinatorial and/or political. In this respect, there is a hugh potential of interesting research problems at the operational level, which could lead to an improved on-line decision support system for KLM's Line Maintenance department.

6.7 Acknowledgements

The author wants to thank Enryk Bakker, Peter Bos, Jos Goedhart, Ben Lammerse, and Klaas de Waal of KLM's maintenance department for their cooperation, and useful suggestions during the development of the decision support system.

Chapter 7

Towards a decision support system for coordinated planning and scheduling of production and maintenance

In this thesis, we have presented a variety of interesting mathematical models, which could be used to assist in, or at least provide insight into the optimization of preventive maintenance policies for complex systems. In each of these models, we haven taken into account possible interactions with production in several dimensions. In this last chapter, we will briefly summarize the ideas and models presented in this thesis. In addition, we will indicate some interesting opportunities for further research in view of a decision support system for coordinated planning and scheduling of production and maintenance.

7.1 Conclusions of this thesis

Let us start with a brief summary of the ideas and models presented in this thesis. As a starting point, we developed a mathematical framework with which the times and/or costs associated with preventive and/or corrective maintenance can be modelled to a proper level of detail. The main underlying observation behind this modelling framework was that most production systems can be decomposed hierarchically into a tree-like structure of set-up activities and components. Since different components may require one or more shared set-up activities, there was a perspective of significant gains if maintenance activities were carried out simultaneously. In chapter 1, we concluded that a clear distinction between static, dynamic grouping and opportunistic grouping strategies was necessary. Moreover, we had to distinghuish between time-based, use-based, condition-based and failure-based maintenance strategies.

Within this modelling framework, the possibilities for static grouping (clustering) of preventive maintenance activities were further exploited in chapters 2 and 3 of this thesis. In chapter 2, we considered a direct clustering problem for multi-setup multi-component production systems with frequency-constrained maintenance jobs. Here, our objective was to subdivide a collection of maintenance jobs into several maintenance packages, such that overall preventive maintenance costs per unit of time were minimized. In chapter 3, we examined a somewhat similar but indirect clustering problem, in which the frequency constraints were replaced by frequency-dependent costs, and each component was maintained preventively at integer multiples of a certain basis interval. This time, our objective was to determine a repetitive maintenance cycle which minimizes the long run average maintenance costs per unit of time. Based upon a series of numerical experiments, we concluded that static maintenance grouping is a powerful instrument to improve efficiency in terms of set-up avoidance.

Subsequently, chapter 4 was concerned with the interval availability distribution of an unreliable production system, which is maintained preventively at regular intervals, and correctively upon failure. Within this setting, we examined the effect of preventive maintenance policies on the guaranteed performance of a production system during a finite period of time, rather than its average performance in the long run. Based upon a variety of numerical experiments, we concluded that the variability of a production system is often a more appropriate performance measure than its availability, which is commonly used in maintenance optimization models. This provided us with a potentially valuable insight. After all, random breakdowns are one of the most disrupte sources of variability in practice, and as such can be reduced by effective maintenance strategies.

In chapter 5, we investigated the potential benefits of building in some flexibility concerning the starting time of preventive maintenance in an operational planning phase. The underlying observation behind this approach was that the initiation of preventive maintenance should be based on the technical state as well as the operating state of a production system, and that the latter is often subject to fluctuations in time. A two-stage maintenance policy was considered, which - in a first stage - used the technical state of the production system to determine a finite interval during which preventive maintenance must be carried out, and - in a second stage - used the operating state of the production system to determine the optimal starting time within that interval. Computational results indicated that significant savings could be obtained in comparison with classical maintenance strategies.

Finally, chapter 6 comprised the results of a case study that was carried out at the Line Maintenance department of KLM Royal Dutch Airlines. This department is

responsible for the inspection, maintenance and repair of aircrafts during their stay at Schiphol Airport, as well as the assignment of aircrafts to flights within KLM's timetable. A decision support system was developed which should eventually assist maintenance managers in determining how many maintenance slots of which type should be available in the timetable, and how many maintenance engineers of which type should be assigned to these slots, in order to comply with the service levels set by higher management. The main complicating factors in this respect were the variation and uncertainty associated with corrective maintenance jobs.

Summarizing, the ideas and models presented in this thesis have addressed a variety of interesting problem areas, in view of the possible interactions between production and maintenance. In our opinion, these problem areas could provide the basis for the design of a decision support system for the coordinated planning and scheduling of production and maintenance. In view of the development of such a decision support system, the main conclusions of this thesis can be summarized as follows:

- in modelling the times and/or costs associated with preventive and/or corrective maintenance, the tree-like structure of set-up activities and components provides a powerful compromise between the theoretical complications and practical limitations of maintenance grouping problems;
- in defining the objective of mathematical models for maintenance optimization, one should remember that the guaranteed capacity of a production system during a finite period of time is often a more appropriate performance measure than its average production capacity in the long run;
- in formulating optimal maintenance strategies for unreliable production systems, there is a perspective of significant reductions in both maintenance times and costs, if some kind of flexibility is build in concerning the starting time of preventive maintenance in an operational planning phase;
- in constructing a schedule of production jobs with intermediate maintenance slots, one should explicitly account for the variations and uncertainties associated with corrective maintenance jobs, in terms of the underlying arrival processes, repair times, and due dates.

In the remainder of this chapter, we will briefly discuss several functionalities of a decision support system for coordinated planning and scheduling of production and maintenance, in view of these and other considerations. Moreover, we will discuss some interesting opportunities for further research as well.

7.2 A framework for design

In order to arrive at a decision support system for coordinated planning and scheduling of production and maintenance, it is in our view essential to maximally exploit the opportunities for static grouping, dynamic grouping and opportunistic grouping. In addition, we should account for time-based, use-based, condition-based and failure-based maintenance policies as well. On the other hand, it is virtually impossible to support the underlying decision problems with mathematical models in each planning phase. In our opinion, preventive maintenance frequencies should be determined at a strategical level, whereas tactical and operational decision making should be supported with relatively simple, and rather straightforward control mechanisms. At the very least, these control mechanisms should facilitate the possibilities for mutual interactions with production planning and scheduling.

In this respect, and in line with our newly developed two-stage maintenance policy, it is completely natural to provide each maintenance activity and/or maintenance package with a release and due date in a tactical and operational planning phase. In our view, this elementary concept is fundamental for the notion of coordinated planning and scheduling of production and maintenance activities. By doing this, preventive maintenance is no longer a prescheduled event for production planning and scheduling, that causes lower machine capacity, but is also a production need that needs to be managed together with production jobs. The other way around, separate planning and scheduling processes for production and maintenance require some compromise to ensure mutual compatibility, thereby involving the risk of sub-optimization.

Although this concept of release and due dates for maintenance activities is common sense in many practical situations, it certainly is an underexposed point of view in existing literature. A typical example of this type was found at the Royal Dutch Airforce, where it is decided at a strategical level that F16's must undergo overhaul maintenance somewhere between 190 and 210 flight hours, but final decisions are made in an operational planning phase. The main underlying justification for the introduction of release and due dates, or so-called **maintenance windows**, is that small deviations from the optimal maintenance interval usually involve low incremental costs, but provide significantly more flexibility in each of the following dimensions:

- workload balancing;
- dynamic and opportunistic grouping;
- coordination with production planning and scheduling.

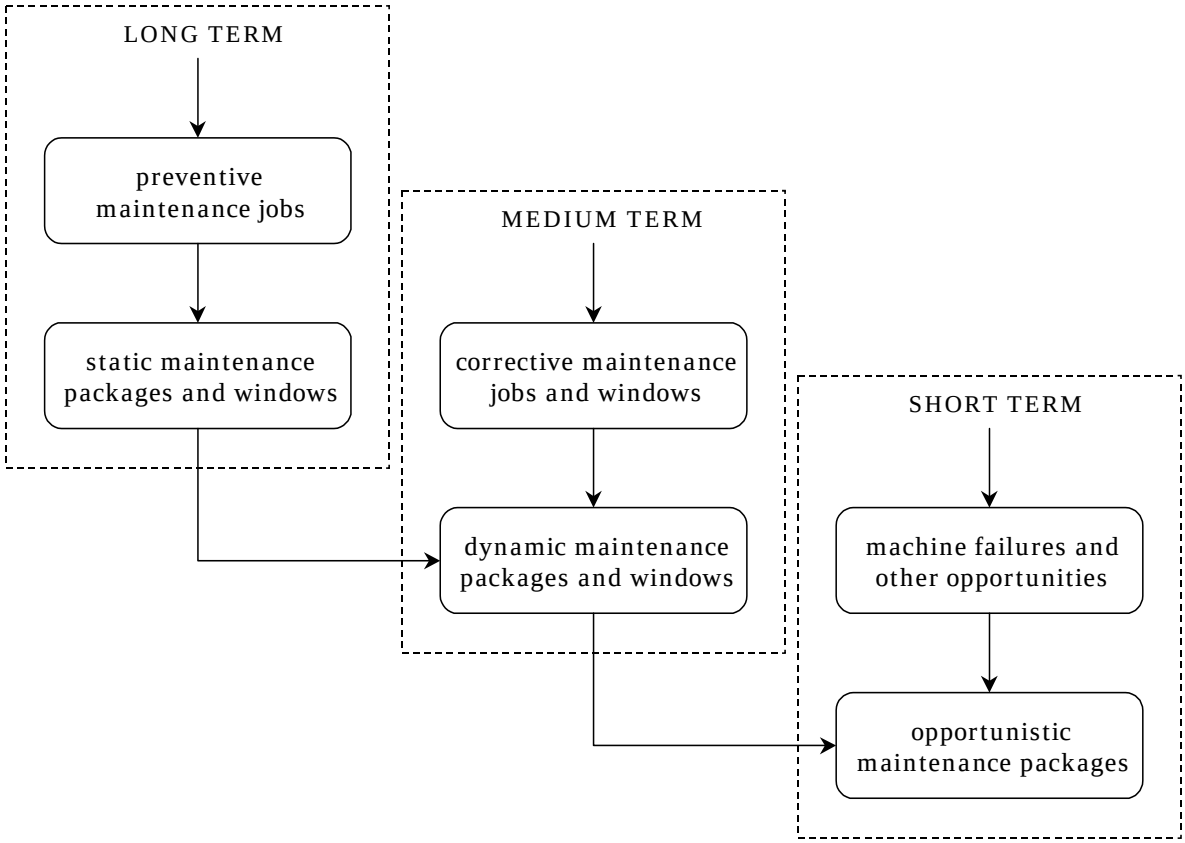


Figure 7.1: Long term, medium term and short term maintenance planning in view of a decision support system for coordinated planning of production and maintenance.

At the very least, a decision support system for coordinated planning and scheduling of production and maintenance should provide the functionalities to define static maintenance packages and maintenance windows in a strategical planning phase. Ideally, it should also be able to evaluate the consequences of user-made decisions in several dimensions (e.g. costs, availability and/or variability), and to suggest alternative solutions on request. Moreover, it should be able to assist maintenance managers in the construction of dynamic and opportunistic maintenance packages in a tactical and operational planning phase, as soon as more detailed information about the technical and operating state of the production equipment becomes available (see Figure 7.1). In this respect, the definition of static maintenance packages and corresponding maintenance windows plays an important, if not crucial role in the context of such a decision support system, since it provides the constraints for further decision support systems and accompanying mathematical models at lower planning levels.

Obviously, the mutual relationships between static, dynamic and opportunistic maintenance packages and maintenance windows at different planning levels can be-

come quite complicated. Therefore, the potential benefits of mathematical decision support models are beyond any doubt. In this thesis, a variety of interesting modelling tools have been presented, each or a combination of which could be very useful within the above-mentioned functionalities of such a decision support system. Nevertheless, there is still a hugh potential of interesting problem areas at each planning level, which have not been addressed at all in this thesis. In the remainder of this chapter, we will briefly discuss some interesting problem areas, and several opportunities for decision support models as well.

7.3 Suggestions for further research

To conclude this thesis, we will elaborate upon some interesting opportunities for further research in view of a decision support system for coordinated planning and scheduling of production and maintenance. To structure this discussion, a clear distinction will be made between long term, medium term and short term maintenance planning, since each of them requires completely different control mechanisms, and accompanying decision support models as well.

7.3.1 Long term maintenance planning

In the long term, a decision support system should assist maintenance managers in the definition of preventive maintenance packages (static grouping) and associated maintenance windows (see Figure 7.2). Ideally, it should also be able to evaluate the consequences of user-made decisions, and to suggest alternative solutions on request. Interesting performance criteria in this respect are e.g. the long run average maintenance costs per unit of time, as well as the availability and variability of the underlying production equipment. In order to arrive at reasonable estimates of such performance measures, it is of crucial importance to have some kind of information with respect to the failure behavior of different components. In the remainder of this chapter, we will assume that such data are available. Further details are skipped, since they are not so relevant for what follows.

In view of long run average performance criteria, it might be optimal to create a few large maintenance packages, in order to minimize the overall times and/or costs associated with preventive and/or corrective maintenance. The other way around, if it is the short term behavior of a production system that counts, one could prefer a large number of relatively small maintenance packages, such that the resulting workload pattern can be evenly distributed over time. The underlying observation behind

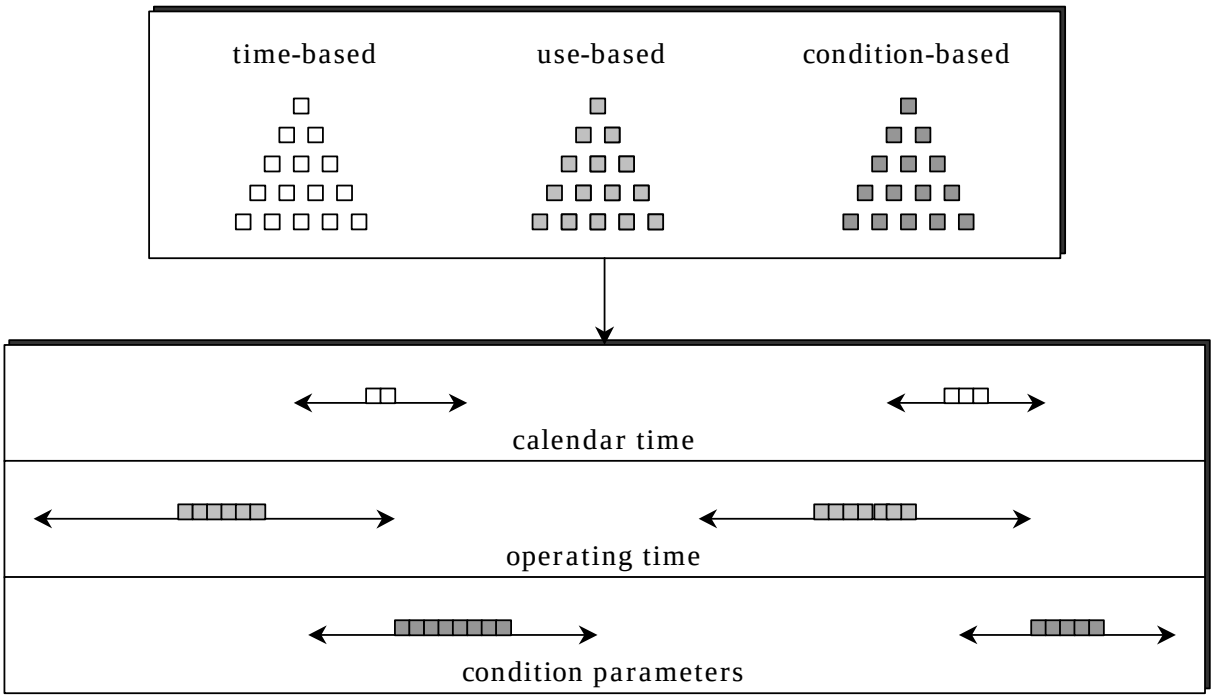


Figure 7.2: Long term preventive maintenance planning: definition of static maintenance packages with time-based, use-based and/or condition-based maintenance windows.

this reasoning is that the variability of a production system does not only relate to the uncertainties associated with corrective maintenance, but also to the variations arising from preventive maintenance activities. Following a similar argument, the construction of maintenance windows for maintenance packages is also a complex issue, with several conflicting objectives as well. In this respect, there is a potential of interesting problem areas at this strategical planning level, which could be supported with mathematical models.

Another important aspect is that maintenance packages should be defined in such a way that they can easily be incorporated in production plans and schedules in a tactical/operational planning phase. In a setting of scheduled production jobs with intermediate maintenance slots, it is therefore important that the time required for static maintenance packages and corrective maintenance jobs, is in agreement with the available maintenance slots in the production schedule. Complicating factors in this respect are the variations and uncertainties associated with corrective maintenance, in terms of the underlying arrival processes, repair times, and due dates. Therefore, some mutual coordination must take place in the construction of production schedules, static maintenance packages, and maintenance windows.

Although each of the models presented in this thesis contains some interesting elements in view of these and other considerations, there is a lack of decision support models which combine these issues into a more comprehensive modelling framework. As an illustrative example, and in view of the ideas and models presented in this thesis, it might be worthwile to examine the interval availability distribution of a multi-setup multi-component production system, in which each component is maintained preventively at an integer multiple of a certain basis interval, and some flexibility is build in concerning the starting time of preventive maintenance as well.

7.3.2 Medium term maintenance planning

In the medium term, a decision support system should assist maintenance managers in the construction of dynamic maintenance packages, each of which is a combination of static maintenance packages and/or corrective maintenance jobs. At this planning level, this means that all use-based and condition-based maintenance windows must first be converted into time-based maintenance windows, in order to be comparable and compatible, and also to facilitate the coordination with production planning and scheduling. If some kind of information is available with respect to the expected utilization of the production equipment in the near future, as well as the expected deterioration processes of the components under consideration, this can usually be done in a rather straightforward manner.

Now that each static maintenance package is provided with a time-based maintenance window, or equivalently with a release and due date, it is the objective of medium term maintenance planning to decide which static maintenance packages and corrective maintenance jobs must be combined with each other, in order to arrive at dynamic maintenance packages with corresponding maintenance windows (see Figure 7.3). A complicating factor in this respect is how to determine the maintenance window for a dynamic maintenance package. If the processing times of the underlying maintenance packages and jobs are relatively small and can as well be neglected, this can be done rather straightforwardly by taking the latest of all release dates, and the earliest of all due dates. In all other cases, the time window of a dynamic maintenance package formally depends on the sequencing of the underlying activities. For example, a worst-case or best-case scenario could be used.

Nevertheless, it is immediately clear from these observations that there is another potential of conflicting objectives at this tactical planning level. Simply stated, the construction of relatively large dynamic maintenance packages may be preferred from an efficiency point of view, but at the same time imply a decrease in the size of

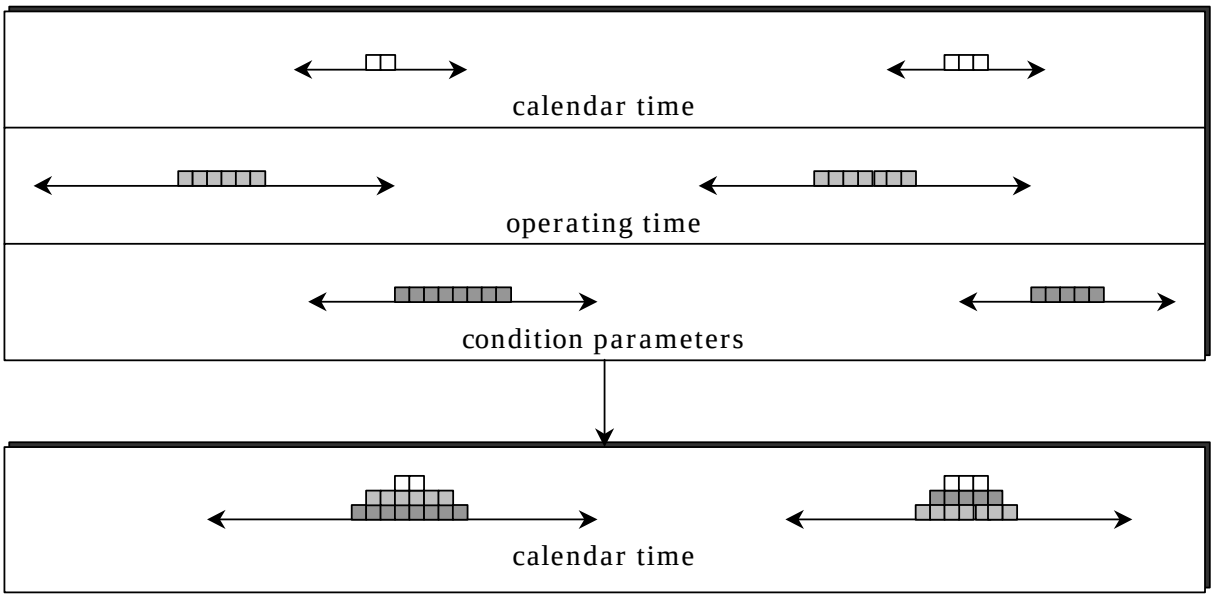


Figure 7.3: Medium term preventive maintenance planning: construction of dynamic maintenance packages with time-based maintenance windows.

the corresponding maintenance windows. On its turn, this might eventually lead to significantly less flexibility in an operational planning phase, and thus lead to unexpected inefficiencies after all. The other way around, creating a large number of relatively small dynamic maintenance packages may provide enough flexibility, but at the same time be inattractive from an efficiency point of view. Once again, finding the right balance between times, costs and/or flexibility is an interesting problem area, which could be supported with mathematical models.

An illustrative example of this type of maintenance planning was presented by Wildeman, Dekker, and Smit (1997), who developed a modelling framework for the dynamic grouping of preventive maintenance activities. Within this modelling framework, the release and due dates for preventive maintenance activities are replaced with a penalty cost for each maintenance activity, which is derived from the discrepancy between its actual and optimal starting time. Nevertheless, the basic underlying concepts are the same, and provide some interesting opportunities for generalizations in several directions. As a starting point, it would be interesting to replace the penalty cost functions with release and due dates for each maintenance activity. Moreover, we could incorporate the hierarchical tree-like structure of multiple interrelated set-up activities and components, in order to arrive at a more realistic cost structure. Finally, it might also be worthwile to build in some additional restrictions with respect to the size of feasible maintenance packages, in view of the possibilities for mutual

coordination with production planning and scheduling in an operational planning phase.

7.3.3 Short term maintenance planning

In the short term, a decision support system should assist maintenance managers in the coordinated scheduling of production jobs and dynamic maintenance packages. At this planning level, production jobs and maintenance packages should be treated on an equal basis, with an open eye for their mutual interactions. In this respect, due date violations for maintenance packages should be handled with the same care as those for production jobs. After all, maintenance should no longer be seen as a necesarry evil, but as a production need that should be managed together with production. Simply stated, it is the objective of short term maintenance planning to arrive at an integrated schedule of production and maintenance activities, in which all release and due dates are satisfied (see Figure 7.4). Of course, it depends on the quality of long term and medium term decision making, whether and up to which degree this objective can be realized in an operational planning phase.

Another important aspect of short term maintenance planning is the optimal use of so-called maintenance opportunities, which may occur due to e.g. idle times, machine failures and/or withdrawn orders. In general, these opportunities cannot be predicted in advance, and are of restricted duration as well. As a result of this, maintenance management often fails to make effective use of them. Therefore, a decision support system should assist maintenance managers in compiling a list of executable dynamic maintenance packages on request, and by setting priorities for each dynamic maintenance package as well. Within this setting, the main questions are how to define the priority of a maintenance package, and how to decide which maintenance packages should be assigned to a maintenance opportunity. Once again, mathematical models might be useful to support such decision making.

An illustrative example of this type of maintenance planning was presented by Dekker and Smeitink (1994), who developed a decision support system for preventive maintenance planning at opportunities of restricted duration. They consider a somewhat similar modelling framework, which does not make use of predefined maintenance windows. As an alternative, they incorporate these release and due dates implicitly in their priority setting procedure. Subsequently, they provide the user with an interactive knapsack scheduling problem, which determines an optimal selection of maintenance packages given the time constraints. Once again, it would be interesting to incorporate our complex structure of multiple interrelated set-up activ-

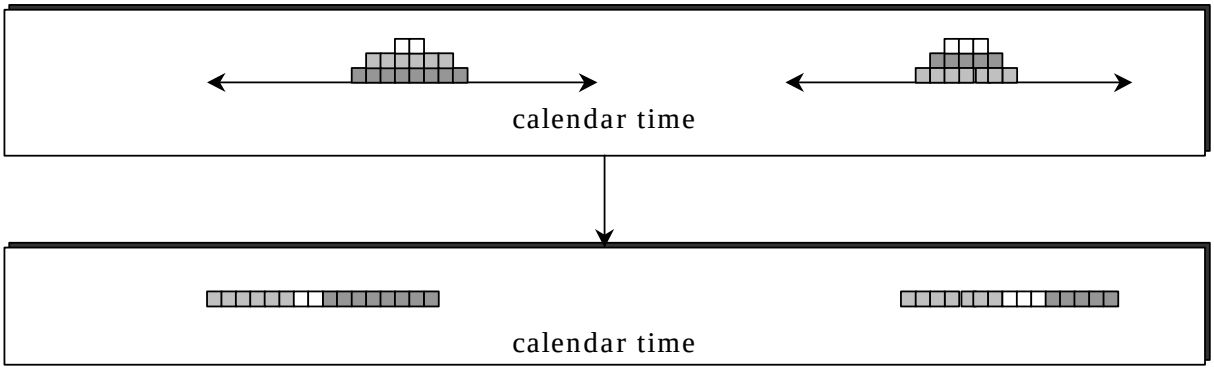


Figure 7.4: Short term preventive maintenance planning: precise scheduling of dynamic maintenance packages within predefined time windows.

ities and components into such a modelling framework. Obviously, this would leave us with a much more complex optimization problem, which has not been addressed so far in existing literature.

7.4 Final remarks

Summarizing, the ideas and models presented in this thesis have covered a variety of interesting problem areas, which are related to the interactions between production and maintenance in several dimensions. As such, they have provided us with useful insights that were previously not available. Nevertheless, there is still a lot of work to be done in order to arrive at an adequate, model-based decision support system for coordinated planning and scheduling of production and maintenance. In this thesis, maintenance has met production, and some cross-fertilisation has taken place. Now they got to know each other, it is time for them to develop a more intense relationship. Of course, we'll stay in touch.

Bibliography

Assaf, D. and J. Shanthikumar (1987). Optimal group maintenance policies with continuous and periodic inspections. *Management Science 33*, 1440–1452.

Aven, T. (1993). On performance measures for multistate monotone systems. *Reliability Engineering and System Safety 41*, 259–266.

Barlow, R. and L. Hunter (1960). Optimum preventive maintenance policies. *Operations Research 8*, 90–100.

Barlow, R., L. Hunter, and F. Proschan (1963). Optimum checking procedures. *Journal of the Society for Industrial and Applied Mathematics 4*, 1078–1095.

Barlow, R. and F. Proschan (1965). *Mathematical Theory of Reliability*. New York: Wiley.

Barron, R. (1996). *Engineering Condition Monitoring : Practice, Methods and Applications*. Harlow: Longman.

Bäckert, W. and D. Rippin (1985). The determination of maintenance strategies for plants subject to breakdown. *Computers and Chemical Engineering 9*, 113–126.

Ben-Daya, M. and M. Hariga (1995). Comparative study of heuristics for the joint replenishment problem. *Omega : the International Journal of Management Science 23*, 341–344.

Berg, M. (1978). General trigger-off replacement procedures for two-unit systems. *Naval Research Logistics Quarterly 25*, 15–29.

Berg, M. (1980). Marginal cost analysis for preventive replacement policies. *European Journal of Operational Research 4*, 136–142.

Berg, M. (1984). A preventive replacement policy for units subject to intermittent demand. *Operations Research 32*, 584–595.

Berg, M. and B. Epstein (1976). A modified block replacement policy. *Naval Research Logistics Quarterly 23*, 15–24.

Berg, M., M. Posner, and H. Zhao (1994). Production-inventory systems with unreliable machines. *Operations Research 42*, 111–118.

Bomberger, E. (1966). A dynamic programming approach to the lot size scheduling problem. *Management Science 12*, 778–784.

Bootsma, P. (1997). *Airline Flight Schedule Development*. Ph. D. thesis, University of Twente, Enschede, The Netherlands.

Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs: Prentice Hall.

Cho, D. and M. Parlar (1991). A survey of maintenance models for multi-unit systems. *European Journal of Operational Research 51*, 1–23.

Christer, A. (1982). Modelling inspection policies for building maintenance. *Journal of the Operational Research Society 33*, 723–732.

Christer, A. and W. Waller (1984). Delay time models for industrial maintenance problems. *European Journal of Operational Research 35*, 401–406.

Cox, D. (1962). *Renewal Theory*. New York: Wiley.

Csenki, A. (1995). An integral equation approach to the interval availability of systems modelled by finite semi-Markov processes. *Reliability Engineering and System Safety 47*, 37–45.

Dagpunar, J. (1982). Formulation of a multi item single supplier inventory problem. *Journal of the Operational Research Society 33*, 285–286.

Dagpunar, J. (1996). A maintenance model with opportunities and interrupt replacement options. *Journal of the Operational Research Society 47*, 1406–1409.

De Koster, M. (1988). *Capacity Oriented Analysis and Design of Production Systems*. Ph. D. thesis, University of Eindhoven, The Netherlands.

De Souza e Silva, E. and H. Gail (1986). Calculating cumulative operational time distributions of repairable computer systems. *IEEE Transactions on Computers 35*, 322–332.

De Souza e Silva, E. and H. Gail (1989). Calculating availability and performability measures of repairable computer system using randomization. *Journal of the Association for Computing Machinery 36*, 171–193.

Dekker, R. (1995). Integrating optimisation, priority setting, planning and combining of maintenance activities. *European Journal of Operational Research 82*, 225–240.

Dekker, R. and M. Dijkstra (1992). Opportunity-based age replacement: exponentially distributed times between opportunities. *Naval Research Logistics 39*, 175–190.

Dekker, R. and E. Smeitink (1991). Opportunity-based block replacement. *European Journal of Operational Research 53*, 46–63.

Dekker, R. and E. Smeitink (1994). Preventive maintenance at opportunities of restricted duration. *Naval Research Logistics 41*, 335–353.

Dekker, R., R. Wildeman, and F. Van der Duyn Schouten (1997). A review of multi-component maintenance models with economic dependence. *Mathematical Methods of Operations Research 45*, 411–435.

Dijkstra, M., L. Kroon, J. Van Nunen, and L. Van Wassenhove (1994). Planning the size and organization of KLM's aircraft maintenance personnel. *Interfaces 24*, 47–58.

Dunn, O. and V. Clark (1974). *Applied Statistics : Analysis of Variance and Regression*. New York: Wiley.

Frenk, H., R. Dekker, and M. Kleijn (1997). A unified treatment of single component replacement models. *Mathematical Methods of Operations Research 45*, 437–454.

Garey, M. and D. Johnson (1979). *Computers and Intractability: a Guide to the Theory of NP-completeness*. New York: Freeman.

Gertsbakh, I. (1972). Optimum choice of preventive maintenance times for a hierarchical system. *Automated Control Computer Sciences 6*, 24–30.

Gertsbakh, I. (1977). *Models of Preventive Maintenance*. Oxford: North-Holland.

Gits, C. (1984). *On the Maintenance Concept for a Technical System: a Framework for Design*. Ph. D. thesis, University of Eindhoven, The Netherlands.

Gits, C. (1987). On the maintenance concept for a technical system III : design framework. *Maintenance Management International 6*, 223–237.

Goyal, A. and A. Tantawi (1988). A measure of guaranteed availability and its numerical evaluation. *IEEE Transactions on Computers 37*, 25–32.

Goyal, S. (1973). Determination of economic packaging frequency of items jointly replenished. *Management Science 20*, 232–235.

Goyal, S. (1974). Determination of optimum packaging frequency of items jointly replenished. *Management Science 21*, 436–443.

Goyal, S. (1982). A note on formulation of the multi-item single supplier inventory problem. *Journal of the Operational Research Society 33*, 287–288.

Goyal, S. and A. Gunasekaran (1992). Determining economic maintenance frequency of a transport fleet. *International Journal of Systems Science 4*, 655–659.

Goyal, S. and M. Kusy (1985). Determining economic maintenance frequencies for a family of machines. *Journal of the Operational Research Society 36*, 1125–1128.

Goyal, S. and A. Satir (1989). Joint replenishment inventory control: deterministic and stochastic models. *European Journal of Operational Research 38*, 2–13.

Groenevelt, H., L. Pintelon, and A. Seidmann (1992). Production batching with machine breakdowns and safety stocks. *Operations Research 40*, 959–971.

Hakes, C. (1991). *Total Quality Management: the Key to Business Improvement*. London: Chapman and Hall.

Hariga, M. (1994). Two new heuristic procedures for the joint replenishment problem. *Journal of the Operational Research Society 45*, 463–471.

Haukaas, H. and T. Aven (1996). Formulae for the down time distribution of a monotone system observed in a time interval. *Reliability Engineering and System Safety 52*, 19–26.

Haurie, A. and P. L'Ecuyer (1982). A stochastical control approach to group preventive replacement in a multicomponent system. *IEEE Transactions on Automatic Control 27*, 387–393.

Heyman, D. and M. Sobel (1982). *Stochastic Models in Operations Research*, Volume 1. New York: McGraw-Hill.

Hopp, W. and M. Spearman (1996). *Factory Physics: Foundations of Manufacturing Management*. Chicago: Irwin.

Jansen, J. and F. Van der Duyn Schouten (1995). Maintenance optimization on parallel production units. *IMA Journal of Mathematics Applied in Business and Industry 6*, 113–134.

Jorgenson, D., J. McCall, and R. Radner (1967). *Optimal Replacement Policy*. Amsterdam: North-Holland.

Kaspi, M. and M. Rosenblatt (1983). An improvement of Silver's algorithm for the joint replenishment problem. *IIE Transactions 15*, 264–269.

Lee, H. and M. Rosenblatt (1987). Simultaneous determination of production cycle and inspection schedules in a production system. *Management Science 33*, 1125–1136.

Lee, H. and M. Rosenblatt (1989). A production and maintenance planning model with restoration cost dependent on detection delay. *IIE Transactions 21*, 368–375.

Lindley, D. (1952). The theory of queues with a single server. *Proceedings of the Cambridge Philosophical Society 48*, 277–289.

Luenberger, D. (1984). *Linear and Nonlinear Programming*. Reading: Addison-Wesley.

McCall, J. (1965). Maintenance policies for stochastically failing equipment: a survey. *Management Science 11*, 493–524.

Meller, R. and D. Kim (1996). The impact of preventive maintenance on system cost and buffer size. *European Journal of Operational Research 95*, 577–591.

Mine, H., H. Kawai, and Y. Fukushima (1981). Preventive maintenance of an intermittently-used system. *IEEE Transactions on Reliability 30*, 391–392.

Moinzadeh, K. and P. Aggarwal (1997). Analysis of a production/inventory system subject to random disruptions. *Management Science 43*, 1577–1588.

Nakajima, S. (1988). *Introduction to TPM: Total Productive Maintenance*. Portland: Productivity Press.

Niebel, B. (1994). *Engineering Maintenance Management* (2nd ed.). New York: Dekker.

Owusu, J. and H. Jessurun (1993). Het ontwerp van een werkstroombeheersingssysteem voor het lijnonderhoud van de KLM-vloot (in Dutch). Technical report, Business Unit Line Maintenance, KLM Royal Dutch Airlines.

Ozekici, S. (1988). Optimal periodic replacement of multicomponent reliability systems. *Operations Research 36*, 542–552.

Pierskalla, W. and J. Voelker (1976). A survey of maintenance models: the control and surveillance of deteriorating systems. *Naval Research Logistics Quarterly 23*, 353–388.

Pintelon, L. and L. Gelders (1992). Maintenance management decision making. *European Journal of Operational Research 58*, 301–317.

Pintelon, L., L. Gelders, and F. Van Puyvelde (1997). *Maintenance Management*. Leuven: Acco.

Ritchken, P. and J. Wilson (1990). (m,T) group maintenance policies. *Management Science 36*, 632–639.

Schäbe, H. (1996). Computing instationary availability for maintained systems. *Microelectronics Reliability 36*, 55–70.

Sculli, D. and A. Suraweera (1979). Tramcar maintenance. *Journal of the Operational Research Society 30*, 809–814.

Sherif, Y. and M. Smith (1981). Optimal maintenance models for systems subject to failure - a review. *Naval Research Logistics Quarterly 28*, 47–74.

Silver, E. (1976). A simple method for determining order quantities in joint replenishments under deterministic demand. *Management Science 22*, 1351–1361.

Silver, E., D. Pyke, and R. Peterson (1998). *Inventory Management and Production Planning and Scheduling* (3rd ed.). New York: Wiley.

Smith, M. (1997a). An approximation of the interval availability distribution. *Probability in the Engineering and Informational Sciences 11*, 451–467.

Smith, M. (1997b). *On the Availability of Failure Prone Systems*. Ph. D. thesis, Erasmus University Rotterdam, The Netherlands.

Srinivasan, M. and H. Lee (1996). Production-inventory systems with preventive maintenance. *IIE Transactions 28*, 879–890.

Takács, L. (1957). On certain sojourn time problems in the theory of stochastic processes. *Acta Mathematica Academiae Scientiarum Hungaricae 8*, 169–191.

Temme, N. (1994). A set of algorithms for the incomplete Gamma functions. *Probability in the Engineering and Informational Sciences 8*, 291–307.

Tijms, H. (1994). *Stochastic Models: an Algorithmic Approach*. Chichester: Wiley.

Trivedi, K. (1982). *Probability & Statistics with Reliability, Queueing, and Computer Science Applications*. Englewood Cliffs: Prentice-Hall.

Valdez-Flores, C. and R. Feldman (1989). A survey of preventive maintenance models for stochastically deteriorating single-unit systems. *Naval Research Logistics 36*, 419–446.

Van der Duyn Schouten, F. and S. Vanneste (1990). Analysis and computation of (n,N)-strategies for maintenance of a two-component system. *European Journal of Operational Research 48*, 260–274.

Van der Duyn Schouten, F. and S. Vanneste (1993). Two simple control policies for a multicomponent maintenance system. *Operations Research 41*, 1125–1136.

Van der Duyn Schouten, F. and S. Vanneste (1995). Maintenance optimization of a production system with buffer capacity. *European Journal of Operational Research 82*, 323–338.

Van der Eijck, M. (1995). Lijnonderhoud gestroomlijnd. Technical report, Business Unit Line Maintenance, KLM Royal Dutch Airlines.

Van der Heijden, M. (1987). Interval availability distribution for a 1-out-of-2 reliability system with repair. *Probability in the Engineering and Informational Sciences 1*, 211–224.

Van der Heijden, M. and A. Schornagel (1988). Interval uneffectiveness distribution for a k-out-of-n multistate reliability system with repair. *European Journal of Operational Research 36*, 66–77.

Van Dijkhuizen, G. (1995). Modelmatige ondersteuning voor het clusteren van onderhoudsactiviteiten. *VLO Magazine: tijdschrift voor logistiek 4*, 2–5.

Van Dijkhuizen, G. (1997). To fly or not to fly: op weg naar een optimale afstemming tussen dienstregeling, onderhoudsvraag en onderhoudsaanbod (in Dutch). Technical report, Business Unit Line Maintenance, KLM Royal Dutch Airlines.

Van Dijkhuizen, G. and M. Van der Heijden (1998). Preventive maintenance and the interval availability of an unreliable production system. Technical report, University of Twente, The Netherlands. Submitted to *Reliability Engineering & System Safety*.

Van Dijkhuizen, G. and A. Van Harten (1997a). Coordinated planning of preventive maintenance in multi-setup multi-component systems. Technical report, University of Twente, The Netherlands. Submitted to *Management Science*.

Van Dijkhuizen, G. and A. Van Harten (1997b). Optimal clustering of frequency-constrained maintenance jobs with shared set-ups. *European Journal of Operational Research 99*, 552–564.

Van Dijkhuizen, G. and A. Van Harten (1998a). Two-stage generalized age maintenance of a queue-like production system. *European Journal of Operational Research 108*, 363–378.

Van Dijkhuizen, G. and A. Van Harten (1998b). Two-stage maintenance of a production system with exponentially distributed on- and off-periods. *International Transactions in Operational Research 5*, 79–85.

Van Eijs, M., R. Heuts, and J. Kleinen (1992). Analysis and comparison of two strategies for multi-item inventory systems with joint replenishment costs. *European Journal of Operational Research 59*, 405–412.

Van Rijn, C. and A. Schornagel (1987). STAMP, a new technique to assess the probability distribution of system effectiveness. *The SRS Quarterly digest, Proceedings of the 17th annual meeting of the Systems Reliability Service 17*, 50–58.

Wartenhorst, P. (1993). *Performance Analysis of Repairable Systems*. Ph. D. thesis, Katholieke Universiteit Brabant.

White, H. and L. Christie (1958). Queuing with preemptive priorities or with breakdown. *Operations Research 6*, 79–95.

Wijngaard, J. (1979). The effect of interstage buffer storage on the output of two unreliable production units in series with different production rates. *AIIE Transactions 11*, 42–47.

Wijnmalen, D. and J. Hontelez (1997). Coordinated condition-based repair strategies for components of a multi-component maintenance system with discounts. *European Journal of Operational Research 98*, 52–63.

Wildeman, R. (1996). *The Art of Grouping Maintenance*. Ph. D. thesis, Erasmus University Rotterdam, The Netherlands.

Wildeman, R., R. Dekker, and A. Smit (1997). A dynamic policy for grouping maintenance activities. *European Journal of Operational Research 99*, 530–551.

Wildeman, R., J. Frenk, and R. Dekker (1997). An efficient optimal solution method for the joint replenishment problem. *European Journal of Operational Research 99*, 433–444.

Zheng, X. (1995). All opportunity-triggered replacement-policy for multiple-unit systems. *IEEE Transactions on Reliability 44*, 648–652.

Summary in Dutch

Nog niet zo lang geleden werd het onderhoud van produktiemiddelen als een noodzakelijk kwaad beschouwd, en waren de meeste onderhoudswerkzaamheden correctief van aard. Tegenwoordig wordt meer waarde gehecht aan de continuïteit van productieprocessen, en wordt het strategisch belang van onderhoud algemeen onderkend. Als gevolg hiervan hebben preventieve onderhoudsconcepten aan populariteit gewonnen. Simpel gezegd zijn deze er op gericht onderdelen te vervangen voordat ze kapot gaan. Teveel preventief onderhoud is daarentegen ook weer niet verstandig. Het vakgebied onderhoudsmanagement houdt zich dan ook hoofdzakelijk bezig met de vraag welke onderhoudswerkzaamheden wanneer en op welke wijze moeten worden uitgevoerd, teneinde de totale onderhoudskosten te minimaliseren. Dit proefschrift vormt hierop geen uitzondering.

Bij het ontwerp van een onderhoudsconcept dienen niet alleen de directe onderhoudskosten, maar ook de aan onderhoud gerelateerde indirecte kosten a.g.v. productiestilstand te worden meegenomen. Hoewel indirecte kosten in het algemeen minder goed zichtbaar, laat staan kwantificeerbaar zijn, vormen ze dikwijls het merendeel van de totale kosten. Zo beschouwd spelen de interacties tussen onderhoud en productie een belangrijker rol dan veelal wordt verondersteld. Naast kostenoverwegingen, kunnen immers ook de beschikbaarheid, de betrouwbaarheid en de beheersbaarheid van de productiemiddelen van doorslaggevende betekenis zijn. Met beheersbaarheid wordt hier bedoeld: de mate waarin voorspeld kan worden wanneer de productiemiddelen beschikbaar (up) danwel niet beschikbaar (down) zijn. Eén van de grootste voordelen van preventief ten opzichte van correctief onderhoud, is dan ook dat de eraan verbonden werkzaamheden ruim van tevoren, en op voor de productie geschikte momenten kunnen worden ingepland.

In dit proefschrift worden een aantal tot de verbeelding sprekende wiskundige modellen ontwikkeld, die ondersteuning kunnen bieden bij het bepalen van de juiste verhouding tussen preventief en correctief onderhoud. In het bijzonder richten we ons hierbij op de zojuist beschreven aspecten, namelijk: kosten, beschikbaarheid, betrouwbaarheid en beheersbaarheid. Om voor de hand liggende redenen, ontwikke-

len we allereerst een krachtig raamwerk waarbinnen de tijd en/of kosten die gepaard gaan met het uitvoeren van preventieve en/of correctieve onderhoudswerkzaamheden tot op een hoog detailniveau kunnen worden gemodelleerd. Hiertoe decomponeren we een productiesysteem op een hierarchische wijze in subsystemen, componenten, onderdelen, etcetera, totdat uiteindelijk een boomstructuur ontstaat met onderling gerelateerde set-up en onderhoudsactiviteiten.

Binnen dit raamwerk ontwikkelen we in hoofdstuk 2 een methode om een verzameling preventieve onderhoudswerkzaamheden onder te verdelen in een aantal onderling onafhankelijke, disjuncte onderhoudspakketten. We beperken ons hierbij tot onderhoudswerkzaamheden waarvan de frequentie reeds bij voorbaat is opgelegd, bijvoorbeeld uit veiligheidsoverwegingen, maar eventueel vaker uitvoeren natuurlijk is toegestaan. Vervolgens proberen we een zodanige onderverdeling te construeren, dat de gemiddelde preventieve onderhoudskosten per tijdseenheid geminimaliseerd worden. We proberen hierbij een balans te vinden tussen de extra kosten die verbonden zijn aan het vaker dan nodig uitvoeren van bepaalde onderhoudswerkzaamheden, en de besparingen in set-up kosten die gepaard gaan met het gelijktijdig uitvoeren ervan.

In hoofdstuk 3 beschouwen we een hieraan gerelateerde, maar fundamenteel verschillende methode voor het op elkaar afstemmen van preventieve onderhoudswerkzaamheden. Ditmaal wordt iedere onderhoudsactiviteit uitgevoerd op een geheeltallig veelvoud van een zeker basisinterval, en worden onderhoudspakketten op een dynamische manier geconstrueerd. Bovendien worden de correctieve onderhoudskosten, die weer afhangen van de frequentie waarmee preventieve onderhoudsactiviteiten worden uitgevoerd, expliciet meegenomen in de modellering. Vervolgens proberen we de hieruit voortvloeiende onderhoudscyclus zodanig in te richten, dat de gemiddelde preventieve en correctieve onderhoudskosten per tijdseenheid geminimaliseerd worden.

In hoofdstuk 4 onderzoeken we de beschikbaarheid en beheersbaarheid van een storingsgevoelig productiesysteem. Klassieke onderhoudsmodellen bepalen doorgaans een optimale onderhoudsfrequentie, door de gemiddelde beschikbaarheid op lange termijn te maximaliseren. We laten zien dat een dergelijke methode tot verre van optimale oplossingen kan leiden, indien we geïnteresseerd zijn in de gegarandeerde beschikbaarheid op korte termijn (beheersbaarheid). Hiertoe ontwikkelen we een wiskundig model waarmee we - bij een gegeven onderhoudsstrategie - kunnen berekenen met welke kans het productiesysteem op zijn minst gedurende een bepaald percentage in een bepaalde periode operationeel zal zijn. Met behulp van een aantal experimenten tonen we vervolgens aan dat de preventieve onderhoudsfrequentie aanzienlijk zal toenemen, indien de korte termijn beheersbaarheid van het productiesysteem van groter belang wordt geacht dan de lange termijn beschikbaarheid.

In hoofdstuk 5 beschouwen we een productiesysteem dat niet volcontinu, maar slechts met tussenpozen in gebruik is. In een dergelijke situatie geldt natuurlijk dat preventief onderhoud bij voorkeur dient plaats te vinden gedurende periodes dat het productiesysteem niet door productie wordt opgeëist, danwel voor onderhoud beschikbaar is. Aangezien zulke informatie omtrent de productiebehoeftes doorgaans pas op korte termijn bekend is, concluderen we dat het wel eens verstandig kan zijn om een zekere mate van speling mee te geven aan het feitelijke tijdstip waarop preventief onderhoud moet worden uitgevoerd. Vervolgens ontwikkelen we een wiskundig model waarmee de optimale preventieve onderhoudsstrategie van dit type kan worden berekend. Aan de hand van een aantal eenvoudige voorbeelden blijkt vervolgens dat deze strategie sterk samenhangt met de karakteristieken van het productieproces.

In hoofdstuk 6 bespreken we de resultaten van een opdracht die is uitgevoerd bij de afdeling Line Maintenance (lijnonderhoud) van de Koninklijke Luchtvaart Maatschappij (KLM). Deze afdeling is verantwoordelijk voor het inspecteren, onderhouden en repareren van vliegtuigen gedurende hun verblijf op de luchthaven Schiphol, alsmede het toewijzen van vliegtuigen aan vluchten in KLM's vluchtschema. Het voornaamste doel van deze opdracht was om kwantitatief inzicht te verschaffen in de onderlinge relaties tussen de dienstregeling, de onderhoudsvraag, en het onderhoudsaanbod. Dit leidde uiteindelijk tot een decision support systeem, waarmee aan de hand van een aantal fundamentele wachtrijmodellen kan worden bepaald hoeveel grondtijd van welk type er in de dienstregeling moet worden opgenomen, en hoeveel personeel van welke type aan deze grondtijden moet worden toegekend. Doelstelling hierbij was het merendeel van de klachten te kunnen verhelpen, met inachtneming van de hiervoor gestelde normen ten aanzien van minimale grondtijd en maximale doorlooptijd.

Tenslotte wordt in hoofdstuk 7 een overzicht gegeven van de in dit proefschrift verworven inzichten, en worden een aantal mogelijkheden voor verder onderzoek in kaart gebracht. Verder wordt in grote lijnen aangegeven hoe de reeds ontwikkelde, en eventueel nog te ontwikkelen modellen, zouden kunnen worden ondergebracht in een decision support systeem voor onderhoudsmanagement, waarin meer rekening wordt gehouden met de eisen en wensen van productie. Ter verduidelijking wordt hiertoe een onderscheid gemaakt in drie aggregatieniveaus, nl. de lange, de middellange, en de korte termijn. Vervolgens wordt globaal aangegeven in welke facetten interacties met productie op deze verschillende aggregatieniveaus naar voren komen. Aan de hand hiervan worden een aantal mogelijke richtingen voor vervolgonderzoek aangedragen.

Curriculum Vitae

Gerhard van Dijkhuizen werd op 22 september 1970 geboren in Amstelveen, als laatste en daarmee jongste telg in een gezin van vijf kinderen, waarvan enkel jongens. Nog geen 18 jaar later ontving hij zijn VWO-diploma aan het Johannes Calvijn Lyceum te Kampen. Vervolgens stortte hij zich vol enthousiasme in een studie Toegepaste Wiskunde aan de Universiteit Twente te Enschede. Hij werd hierbij vergezeld door een drietal klasgenoten, waaronder zijn iets oudere tweelingbroer Johan. In augustus 1993 rondde hij zijn studie met succes af middels een praktische afstudeeropdracht bij PTT Research te Leidschendam, waarin hij meewerkte aan de ontwikkeling van het toen nog in de kinderschoenen staande GSM netwerk voor mobiele telefonie.

Nog niet uitgekeken op het studentenleven, begon hij in 1994 als Assistent in Opleiding (AIO) aan zijn promotieonderzoek bij de vakgroep Operationele Methoden en Systeemtheorie van de faculteit Technologie & Management, eveneens aan de Universiteit Twente. In de daaropvolgende jaren verrichte hij onderzoek op het gebied van "integratie van produktiebesturing en onderhoudsmanagement", waarvan het resultaat momenteel voor u ligt. De resultaten hiervan leidden uiteindelijk tot een handvol wetenschappelijke publicaties in internationale tijdschriften, en werden gepresenteerd op een viertal conferenties in achtereenvolgens Manchester, Antalya, Vancouver en Brussel.

Ter afronding van zijn promotie-onderzoek, verrichte hij in 1997 een opdracht voor de afdeling Line Maintenance van de Koninklijke Luchtvaart Maatschappij (KLM) op Schiphol Airport. Hier ontwikkelde hij een decision support systeem, waarmee de onderhoudsvriendelijkheid van een (concept) dienstregeling kan worden ingeschat. Niet lang daarna werd hij aangesteld als Universitair Docent (UD) aan de vakgroep waar hij reeds werkzaam was. In deze functie verzorgt hij sindsdien onderwijs en onderzoek in uiteenlopende disciplines, waaronder operationele research, wachttijdtheorie, onderhoudsmanagement en interne logistiek.