Operations Research Models for Scheduling Railway Infrastructure Maintenance

ISBN: 978 90 361 0126 4

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **456** of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Operations Research Models for Scheduling Railway Infrastructure Maintenance

Besliskundige modellen voor onderhoudsplanning van de spoorweginfrastructuur

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. S.W.J. Lamberts

and in accordance with the decision of the Doctorate Board.

The public defense shall be held on

Thursday 4 June 2009 at 13:30 hours

by

Gabriella Budai–Balke
born at Reghin, Romania.

ERASMUS UNIVERSITEIT ROTTERDAM

Doctoral Committee

Promotor:	Prof.dr.ir. R. Dekker
Other members:	Prof.dr. J. Vatn
	Prof.dr. L.G. Kroon
	Prof.dr.ir. U. Kaymak

Acknowledgements

Before summer 2000 I never thought or dreamt that I would end up living and doing a Ph.D. in a foreign country 2000 kilometers away from my home town. But God's plans are often different from our plans. In that summer I got to hear from Csaba and Katalin Boer that Rommert Dekker had a Ph.D. project available in maintenance optimization. Although I was new to this subject, I decided to apply for this position since it would give me a possibility to do research in applied mathematics. After a couple of email exchanges with Rommert my application was accepted and consequently I came to Rotterdam to join first the MPhil programme at the Tinbergen Institute and later, in 2002, the Ph.D. programme that resulted in this thesis. Herewith, I would like to extend my gratitude to several persons who helped me to complete this research project.

First and foremost I would like to thank my supervisor Rommert Dekker for offering me the opportunity to do a Ph.D. at Erasmus University. I am greatly indebted to him for his continuing confidence in me as well as for giving me constant support throughout these years. He has been involved in all studies included in this book and I learnt many things from his rich academic and practical experience.

Next, I would like to thank the members of my Ph.D. doctoral committee for reading this thesis and especially to Jørn Vatn, Leo Kroon and Uzay Kaymak for their constructive comments on an earlier draft of this work.

I am grateful to Gábor Maróti for his immense help and support during my Ph.D. trajectory. I greatly appreciated working together with him on Chapter 6 and I enjoyed the discussions we had about mathematics, programming and life in general. Without his support this thesis would not be the same. A special thanks goes to Dennis Huisman and to Leo Kroon for co-authoring Chapter 6. Furthermore, Dennis has also been involved in the project leading to Chapter 4. The discussions with him helped me to understand better how to deal with modelling and all kinds of railway related issues. I would also like to express my gratitude to Robin Nicolai not only for co-authoring Chapter 2, but also for his friendship, never-ending encouragement and interest in my work. Robin, it was a pleasure to attend courses, workshops and conferences with you and to work together on some papers and in the Dynaform project.

In addition, I thank Uzay Kaymak for working together on the study on which Chapter 5 is based and for sharing his experience on genetic and memetic algorithms. Moreover, I would like to take the opportunity here to thank the two Master students: Joel van 't Wout and Diederik de Ree for their work in the preliminary phase of that study.

I greatly appreciate the financial support of the Econometric Institute and Tinbergen Institute, which enabled me to present my work at various conferences abroad. In addition I would like to thank Tineke Kurtz and Carien de Ruiter for their help with the formalities that had to be arranged when I came to the Netherlands, a process that was then more cumbersome than now, as Romania was at that time not yet a member of the EU. Moreover, I want to thank Elli Hoek van Dijke, Anneke Kop and Marjon van Hees - Gouweleeuw for always being very helpful.

Getting through my Ph.D. trajectory would not have been possible without the support and the encouragements of my friends. I am grateful to Ildikó and Marc Meertens, Katalin and Csaba Boer, Emőke and Bart Oldenkamp, Andrea and Szabolcs Boros and Monica Popa for their friendship and the pleasant time spent together. In this line I would also like to mention several of my old colleagues from the Erasmus University, Marisa de Brito, Eric Porras, Haris Munandar and Sandra Phlippen.

In the past two years, it has been a real challenge for me to combine an almost full-time position at ORTEC with finalising this thesis. I am very thankful for the support of my colleagues from the TNT department, especially to Inge Baan, Arjen van de Wetering and Arjen Rietveld. Many thanks for your help, for showing continuous interest in my research work and for stimulating me to complete this dissertation.

For me above all my professional achievements, living in the Netherlands has come with a wonderful gift as I met here my husband and his loving family to whom I am very grateful. Special thanks go to my dear parents, to Barni and Edithke for their love, for encouraging and supporting me during difficult periods and for never stopping believing in me. *Kedves Anyu es Apu, hálás köszönet mindazért a sok jóért amit értem tettetek. Köszönöm, hogy mindig támogattatok és hittetek bennem.*

Finally, the most important "thank you" goes to my dear husband, Bert. Thank you for your love, for your endless patience, for comforting and encouraging me during the challenging periods. Without your support I would not have been able to complete this thesis.

Gabriella Budai–Balke
Ankeveen, April 2009

Contents

Chapter 1

Introduction

In the last decades, maintenance of technical systems has become increasingly important in many industries. Such systems are transport systems (rail, bus, airplane), civil engineering systems (roads, buildings, bridges), health care (hospitals), communication systems, manufacturing plants, *etc.* Failures of these systems may cause expensive production losses and can have a negative effect on the environment and safety. For example, a train accident, crash of an airplane, collapse of a building, computer breakdown in an hospital, may cause large societal costs. Substantial environmental damage may result from a failure in the chemical industry and an unanticipated breakdown of a production system may cause large economic losses.

By corrective maintenance, including repairs and replacements of the failed components, the failed system can be restored to an operational state. However, by preventive maintenance actions, including condition monitoring, inspections, small repairs, these failures can be prevented or their consequences can be reduced, but unfortunately the failures can never be totally eliminated. As Kobbacy and Murthy (2007) remark, over the last 50 years the approach to maintenance has been significantly changed. While over a hundred years ago, the focus was mostly on corrective maintenance, after the Second World War preventive maintenance gets increasingly more attention.

The costs for maintenance has increased over the years as well. For example, in the Australian freight operations the maintenance costs represent 25–35% of total train operating costs (Higgins (1998)), the cost of maintenance in a highly mechanized mine can be 40–60% of the operating cost (Campbell (1995)) and the maintenance spending in the UK's manufacturing industry ranges from 12 to 23% of the total factory operating costs (Cross (1988)). These cost increases are likely to continue as the infrastructure and other facilities are growing steadily and the manpower costs are increasing as well. This shows that improvement of maintenance management has become an economic and social necessity. Maintenance management systems are often used in organizations, where large-scale preventive maintenance schemes are being set up to reduce the probability of failure.

However, since preventive maintenance is not for free, a good balance should be made between its costs and benefits. Beside the budget constraints, the maintenance engineers are also confronted with questions as *When to perform planned maintenance actions and on which components of the system? What are the (production) losses for shutting down the system for a given time period for inspection, repairs or replacements? How can one make the best use of the downtime once the system is shut down for maintenance? Can I restart after maintenance the operation of my system without any problems?*

This study is an attempt to assist the maintenance managers to answer these questions through reviewing the literature on maintenance optimization, developing useful methods and improving the existing methods for scheduling preventive maintenance works. This first chapter starts with a general description of maintenance scheduling, followed by a section about maintenance scheduling in the railway infrastructures. In Section 1.3 we briefly describe the effect of rail infrastructure maintenance on the train operation. The thesis objectives are given in Section 1.4. Finally, we give an outline of the remainder of this thesis in Section 1.5.

1.1 Maintenance scheduling in general

Maintenance includes all actions necessary for retaining a system or an item in, or restoring it to, a state in which it can perform its required function (British Standard (1984)). All users would like their system to function as long as possible or, at least as long as it is needed. Therefore, it is necessary to maintain the system's functionality by performing appropriate maintenance tasks. Thus, a maintenance task can be defined as a set of activities that need to be performed, in a specified manner, in order to maintain the functionality of the item or system.

Maintenance tasks can be divided into the following three main categories:

- corrective maintenance,
- preventive maintenance,
- predictive maintenance.

Corrective maintenance (CM), also known as breakdown maintenance, is performed to restore a failed or malfunctioned item or systems. This is a reactive approach to maintenance because the action is triggered by an unscheduled event, such as failure of an item. With this kind of maintenance policy, the maintenance related costs are usually high due several reasons. First of all, restoring the item or system mostly has to be done urgently, thus planning the manpower and spare parts is extremely difficult. Secondly, the failure of an item might cause a large amount of consequential damage to other items in the system too. Finally, there are high safety/health dangers caused by the failure and the costs of downtime and penalty associated with the lost production is mostly huge. An example of the challenges posed by corrective maintenance can be seen with the incident that happened near Amsterdam Central Station on November 2008 (ProRail homepage (2009)). The derailment of a cargo train caused extensive damage to the rails as well as to the switches over a length of 500 meters, resulting in high repair/replacement costs on the infrastructure and in the closure of one of the most heavily used railway tracks in the Netherlands for about a week. Although lack of maintenance of the railway track was in itself not the primary cause of the accident, the fact that the track could not be used meant a huge disruption of the train traffic near Amsterdam.

Preventive maintenance (PM) is used to minimize the disadvantages of the CM by reducing the probability of occurrence of failure, preventing sudden failures and discovering hidden failure. PM includes preplanned and scheduled adjustments, major overhauls, replacements, renewals and inspections. Some of these activities will result in system downtime, whereas others can be done while the system is in operation. A big advantage of PM is that it can be planned ahead and performed when it is convenient. This is very important when work preparation is necessary, so for example new components can be ordered in time and also enough maintenance crew can be available at the planned maintenance execution times.

Conditional (predictive) maintenance tasks often refer to condition-monitoring preventive maintenance tasks where direct monitoring methods are used to determine the exact status of the items, for predicting possible degradations and for discovering those areas where maintenance is needed. The objective is to predict the time that failures will occur and to take actions based on the predictions.

Nowadays, the interest for using conditional maintenance tasks has increased because of the safety requirements and a need to reduce the maintenance costs. Waiting until a component fails may maximise the life of that component, but its failure may cause significant damages to other parts of the system. Moreover, it will cause a disruption of the whole operation. But having a complicated system (*e.g.* railway system) with many components, means that it is very difficult (or even impossible) to monitor every one of them and to keep all the information in a database. Therefore, conditional maintenance tasks may not be usable for complex systems. Scheduled preventive maintenance, on the other hand, may be very convenient, but is likely to result in more maintenance than what is strictly needed, because some parts will be replaced when they have passed only a fraction of their expected life. The maintenance cost might be large as well, due to frequently replaced spare parts and labour. However, one way of reducing preventive maintenance costs is to combine the executions of maintenance activities (see *e.g.* Dekker (1995), Van Dijkhuizen and van Harten (1997), Van Dijkhuizen (2000)). In many cases preparatory work, such as shutting down a unit, scaffolding, traveling of the maintenance crew and machines, has to take place before maintenance can be done. Combining activities allows savings on this preparatory work. However, grouping mostly implies that one deviates from the originally planned execution moments and thus some maintenance actions are more often performed than originally planned. This involves costs as well. Now the following questions arise: *How much can we save by combining as much as possible the maintenance activities in comparison with planning each maintenance work separately? How to decide when to combine works and when to plan them separately?* In Chapters 4 and 5 of this thesis we answer these questions. Moreover, we will also take into account the situation when several maintenance works are not allowed to be combined with each other.

In the literature grouping of maintenance activities is modeled on the long term (stationary models) and on the short term (dynamic models). Some important references are Cho and Parlar (1991), Wildeman *et al.* (1997), Dekker *et al.* (1997). In the stationary models a long term stable situation is assumed and mostly these models assume an infinite planning horizon. These kind of models provide static rules for maintenance, which do not change over the planning horizon, so several maintenance activities will always be carried out at the same time until the end of horizon. These models determine, for example, long term maintenance frequencies for groups of related activities. In practice, however, planning horizons are usually finite since information is only available on the short term and even a small modification of the system can change the problem completely. Thus, in the dynamic models, short term information such as a varying deterioration of components or unexpected opportunities can be taken into account. These models generate dynamic decisions that may change over the planning horizon. In practice we often see that because components have mostly longer lifetime than the length of the horizon, a finite horizon is applied in a rolling horizon approach. Thus, the decisions in the finite horizon are implemented, and subsequently a new horizon starts. These type of optimization problems are much more complex, therefore new research is needed.

In Chapters 4 and 5 we consider dynamic grouping of the maintenance activities for a finite planning horizon. In these two chapters our objective is to develop maintenance optimization models for determining the timing of the preventive maintenance activities. The aim of these mathematical models is to find the optimum balance between the costs of deviating from the optimal preventive maintenance schedule for individual maintenance activities and the benefits of combining different maintenance activities (*e.g.* savings on costs), while taking all kinds of constraints into account. Here we also look at the possibilities to combine preventive maintenance activities with plannable corrective maintenance activities, given that repair/replacements of the failed units can be postponed. Moreover, in the mathematical models developed in Chapter 5 the failure of a particular unit in the system is used as an opportunity for planned maintenance on the other units. Thus, the benefit of combining planned maintenance activities with unplanned activities is tackled. Examples of maintenance optimization models where maintenance is performed at randomly occurring opportunities can be found in Özekici (1988), Dekker and Smeitink (1991), Dekker and Dijkstra (1992), Dekker and Smeitink (1994), Dekker and van Rijn (1996), Wijnmalen and Hontelez (1997), *etc.*

In general the maintenance models can be classified as either deterministic or probabilistic. Deterministic problems are those in which the timing and outcome of the maintenance and replacement actions are assumed to be known with certainty. The probabilistic problems are those where the timing and outcome of the maintenance and replacement actions depend on chance. In this thesis we do not deal with probabilistic problems, such as failure modeling. We assume that the randomness has been already covered in an earlier stage of the optimization process and the type and frequency of the preventive maintenance actions is known. Instead, in this thesis we optimize the executions of these maintenance actions. In determining when to perform a maintenance activity we are interested in the sequence of times at which the maintenance activity should take place such that the overall cost is minimized.

In the past years much attention has been paid to the need for improving the quality of modes in the European infrastructure network. Such modes can be main ports (international airports and harbors), but it seems to be more important to focus the attention on major metropolitan areas in Europe, to European infrastructure systems, namely to public transport. The statistics have shown that the demand for railway transport has increased considerably in the past few years. The intensity with which the existing railway network is being used also creates an increasing need for efficient planning of the preventive maintenance works on the railway infrastructure.

Scheduling of the preventive maintenance works on the railway infrastructure is very difficult, since there are many constraints to be considered. Firstly, carrying out maintenance on the rail infrastructure usually involves many disturbances for the travelers (*e.g.* delays, canceled trains), and vice versa, the train operation restricts the length and the frequency of the infrastructure possession.[1] Moreover, in the last years due to a couple of severe accidents the safety regulations for the track workers became very strict. Thus, in some countries no train operation is allowed during maintenance work. Furthermore, the railway infrastructure maintenance costs have increased substantially in the past years. For example, in the Netherlands the maintenance costs increased by 40% between 1994 and 2001 and in 2006 the preventive maintenance costs were approximately 257 million euro (Meeus and Staal (2007)).

Since the railway infrastructure maintenance is very difficult to plan and involves high costs, there is a need for developing operations research tools, which help the mainte-

[1] When a section of track is required for maintenance and it is therefore blocked from train traffic, it is handed over by the operators to the engineers, who take "possession" of the track. When the track is returned to the operators, the engineers "give up possession".

nance planners to come up with optimal maintenance plans. The literature review in Chapter 3 shows that there are not many articles that treat this issue, yet it proves to be an interesting research subject and the results might help to improve the quality of the rail infrastructure. This challenged us to focus in this thesis mostly on railway infrastructure maintenance, however the mathematical models and techniques developed for scheduling railway maintenance can also be applied for maintenance scheduling in other public/private sectors as well.

The term infrastructure covers all the assets that are used for train operation (Improverail (2002)). These assets are:

- Tracks (rails, sleepers, fastenings, ballast), switches and crossings,
- Bridges and tunnels,
- Energy supply installations (catenary systems),
- Safety and telecommunication equipment (signaling systems).

In the next section some maintenance issues related to the Dutch railway network will be discussed, highlighting the special maintenance tasks. Furthermore, we describe the way how the railway infrastructure maintenance planning is currently done in the Netherlands.

1.2 Scheduling of the railway infrastructure maintenance in the Netherlands

In 2008, the railway network in the Netherlands consisted of about 6800 railway kilometers, 5100 bridges and tunnels, 7488 switches and 386 stations (NS homepage (2009) and ProRail homepage (2009)). It handled about 5900 passenger trains and 400 cargo trains per day. Furthermore, every day about 1.2 million passengers and 100,000 ton of cargo were transported. Statistics show that the Dutch railway network is one of the busiest networks in Europe. The network is operated mainly by the Dutch railway company: Netherlands Railways (Nederlandse Spoorwegen, NS) and it is fully financed and owned by the Dutch government. NS manages the rolling stock and the train personnel and *ProRail* is responsible for the availability, reliability, safety and quality of the rail infrastructure. Moreover, ProRail takes also care of new constructions and extensions, such as the Betuweroute and High Speed Line South (for more information about these projects we refer to Zoeteman and Braaksma (2001) and http://en.betuweroute.nl/). The Netherlands is one of the few countries where all the maintenance works, replacements and renewal projects on the railway infrastructure are outsourced by so-called "output process contracts" to three maintenance contractors: Strukton Rail (50% market share), VolkerRail (30% market share) and BAM (20% market share) (see ProRail homepage (2009)). In these contracts a precise description is given of what is expected with respect to quality. Contractors are responsible to achieve this quality. ProRail does not tell the contractors how to carry out the maintenance, but evaluates the quality by using several modern measuring instruments, and by monitoring and analyzing the disturbances (Improverail (2002)).

For a long time, it has been government policy to reduce the road congestion, which has severe economic and ecologic damage. In the Netherlands for example, the government is taking measures to stimulate passenger and freight transport by train as a viable alternative to road transport. To some extent, these measures are successful since in 2007, the Netherlands Railways reports (NS homepage (2009)) an increase of 12% in passenger kilometers in comparison with 2003. Furthermore, the freight transport has increased considerably in the past years as well. In order to satisfy the increasing demand for rail transport more trains are needed and more frequent service, especially in the rush hours. In some cases new infrastructure is necessary. Furthermore, a longer operating time, and safe and comfortable travel is desired. Due to the increase in the number of trains and traffic load the infrastructure deterioration is growing. Accordingly, more maintenance and renewal work is needed and thus the infrastructure possession time for these maintenance activities increases as well. However, more infrastructure possession time decreases the available train-paths and this has as consequence that the infrastructure may not be able to satisfy the demand. In order to solve this conflict, the infrastructure management should come up with clever policies for making train-paths available as required by the market at an acceptable quality level and taking the possession times for maintenance and renewals also into account. Figure 1.1, that is taken from Improverail (2002), shows the conflict situation that the rail infrastructure managements has to deal with.

The duration of the possessions and their effect on passenger train operation varies widely across European railway companies. It depends, for example, on how busy the railway network is during the day, on the frequency of the passenger and freight trains during the night, on whether alternative transportation modes are allowed to substitute the train operation, on the safety requirements of the maintenance crew, *etc.* However, generally speaking the possessions can be divided into three main categories (see Office of the Rail Regulator (2001)):

- overnight possessions
- Sunday and full weekend possessions
- daytime possessions.

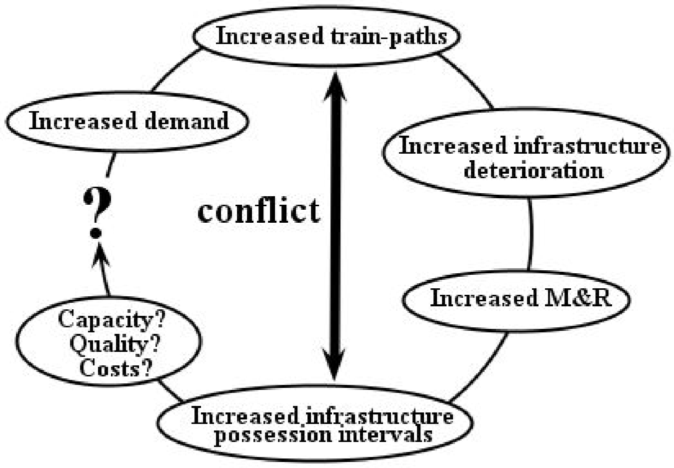

Figure 1.1: Main issues in infrastructure management (from Improverail (2002))

Overnight possessions are taken after the last scheduled passenger train at night and cleared before the first train of the following day. This period is mostly only a few hours long. During this period the maintenance crew and equipment should be moved to the place where maintenance work will be carried out and be removed at the end from there. Thus, the time left for productive work can be extremely limited. The net effect is that overall efficiency is reduced and a large number of possessions is required.

Sunday and weekend possessions affect train operation during a complete weekend or on a Sunday. For such possessions re-scheduling of the train operation is needed. Trains would operate during the possession according to a modified timetable, since some train services might be canceled or trains might be rerouted via adjacent tracks or via longer/shorter routes. Usually, if train services are suspended, then for the travelers alternative transport means are arranged, such as busses, taxis, *etc.* These type of possessions are much more efficient, since the maintenance crew can work continuously for a couple of hours or days. Moreover, during Sunday or weekend the number of passengers is lower, in comparison with the weekdays, thus the nuisance for the passengers is reduced, too.

Possessions during daytime usually involve major disruptions to passenger train services (cargo trains can be rerouted easier), unless the possession is taken during low traffic periods (*e.g.* holidays) or on the tracks that are used occasionally, as it is the case in Australia and in some European countries. In some countries possessions during daytime are hardly ever allowed (*e.g.* the Netherlands). This has several reasons, such as high costs due to the nuisance for the passengers, densely used network during daytime and increased danger for accidents with the maintenance workers. However, in the Netherlands a new development for maintaining the tracks during train operation has recently been tested with success (see VolkerRail homepage (2009)). Using this innovative development, that is called *mobile workshop*, the maintenance crew can work faster, safer and more comfortable on the track while on the adjacent tracks passenger/freight trains are running. In this way the disruption of the train traffic is reduced substantially. The mobile workshop consists of a 25 meter long wagon, without a floor, that is equipped with water, electricity and all of the tools and spare parts that track workers need. This wagon can be quickly and efficiently driven on the track to the work site under his own power or pulled by a locomotive.

Anyhow, we can conclude that the maintenance productivity will be much higher if one can work during a longer period. The other advantage of having longer possessions is that multiple works can be carried out at the same time. Thus different preventive maintenance works can be combined with each other or with deferrable corrective works in order to reduce track possession time or to save execution costs. Usually, the possession planning is done long beforehand, *e.g.* more than a year beforehand in the Netherlands.

In the next section we present the different type of railway infrastructure maintenance works that we consider in this thesis.

Type of railway infrastructure maintenance

The preventive railway infrastructure maintenance can be divided into:

- Routine (spot) maintenance works
- Project (systematic maintenance)

Routine (spot) maintenance works consist of inspections and small repairs of the local irregularities carried out manually or using small machines. These are jobs that do not take much time to be performed and are done frequently from once per month to once in a year. For example: switch inspections, switch lubrication, maintenance at level crossings, rectifying track gauge, tamping using vibrating compactors or tamping tines[2], *etc.* (Esveld (2001)) Inspections of the track are generally performed at regular intervals. The purpose of inspections is to determine whether the condition of the track components is satisfactory or unsatisfactory and hence whether further action is required. We mention two types of inspections, namely visual inspection and condition measurements.

The frequency of the visual inspection varies depending on speed limit and daily train tonnage from a few weeks on the most important lines to once a month on the least important lines (see Esveld (2001)). In general, the visual inspections do not need track possession. Extra inspections are necessary in special cases, such as very hot or very cold weather. The condition measurements, namely ultrasonic rail inspections (with hand equipment or ultrasonic train) are performed in order to check rails for internal defects,

[2] A detailed description of the terminology is given in Appendix A.

rail profile, rail track geometry, rail surface, to carry out measurements on switches. A very important advantage of ultrasonic inspection is that defects can be detected at an early stage, so the repairs can be scheduled on time. A well-known firm concerned with inspection and data analysis is Eurailscout Inspection and Analysis. They also apply video inspections of track surrounding, ballast structure, overhead wire, *etc.* Each part of the track in the Netherlands is checked 4 times a year using the Eurailscout ultrasonic inspection car Strukton Rail homepage (2009).

Project (systematic maintenance) include large amount of work that necessitate separate planning. These activities are carried out with heavy track maintenance machines (*e.g.* tamping machines, ballast regulators, rail-grinding machines, ballast cleaners) approximately once per a couple of years. For example, tamping[2] is done on average every 4 years, grinding every 3 years.

It is difficult to really draw the line between routine maintenance and projects. Each of them has different planning characteristics, such as frequency, duration, *etc.* Furthermore, if you make a planning for a finite horizon, then routine works will be executed several times and scheduling one execution will influence the other executions, while projects can be scheduled independently of their possible future executions.

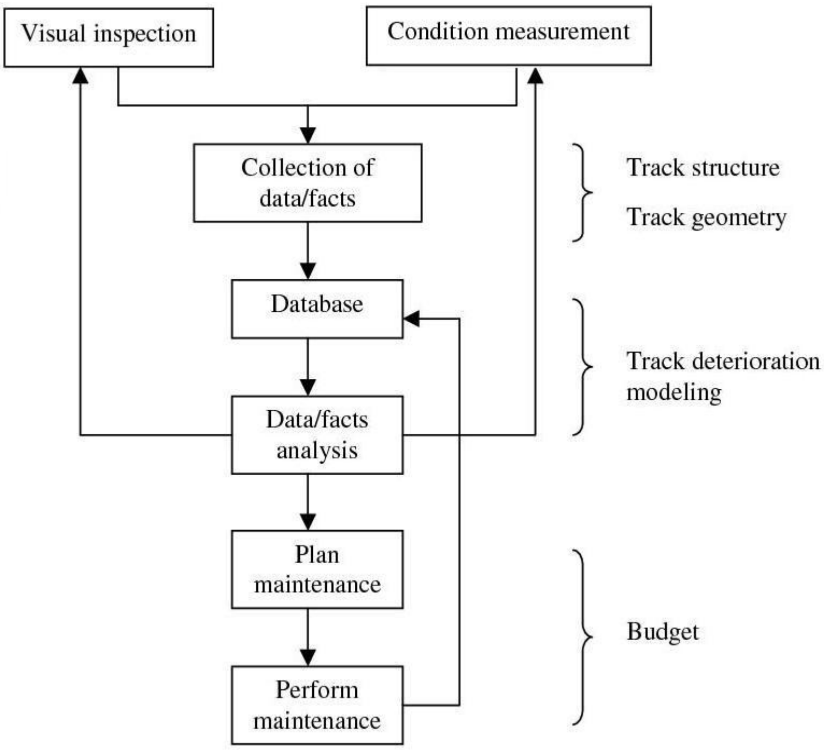

Figure 1.2: An overview of the maintenance planning

In Figure 1.2 we present an overview of the maintenance planning process. After either visual or ultrasonic inspections are effectuated, data and facts related to the track are collected and introduced in the database. Analyzing those data, based on track deterioration models, it is decided whether maintenance work is needed or new inspections need to be planned. If the rail segment (or other track components) has to be maintained and there is enough budget for it, then the maintenance work is planned and carried out. Finally, after the evaluation of the work, the new characteristics of the track are introduced in the database.

There are also other type of works on the railway infrastructure. These are: renewals and new constructions. *Renewal* is done for safety reasons or when the maintenance of different track components is becoming too expensive. Often renewal is included in maintenance, but here we will mention it separately, because that is the practice in the railways. We can mention the following track renewal activities: sleepers, rail, ballast renewal and complete or partial renewal of the switches. *New building* includes all activities that are intended to construct completely new tracks, tunnels, bridges, stations, *etc.* throughout the country.

According to Esveld (2001), the annual maintenance on the Dutch railway network contains renewal of about 140 km of main line and 40 km of secondary tracks, 1000 km of mechanical tamping, 60 km of ballast cleaning, 10 km of corrective grinding and renewal of 250 switches. In addition to this, the track requires spot maintenance on a daily basis. With respect to the costs, Zoeteman (2007) reports that tracks and switches consume more than 50% of total maintenance costs and 75% of renewal costs due to their high usage, relatively rapid deterioration pattern and high installation cost. Tracks and switches are, in fact, the most important equipments of the railway infrastructure, since they are often sources of failure and traffic disruption.

Due to very high maintenance costs, it is important that the track maintenance works are scheduled in an effective and efficient manner. This includes short term planning, such as daily scheduling of the activities, as well as the medium to long term planning. In the next section we give a detailed description of how the railway infrastructure maintenance planning is currently done in the Netherlands.

Current railway infrastructure maintenance planning in the Netherlands

Due to several severe accidents with track workers (*e.g.* in 1995 in Mook[3]), the Dutch government decided that track workers should be protected while working on the rail infrastructure. Thus a new policy for guaranteeing the safety during maintenance works

[3] www.trouw.nl, 10th of June 1995, only in Dutch

was developed (Den Hertog *et al.* (2005)). According to this approach, called "workcrew equals train", if there are track workers in a specific part of the corridor, that part should be taken out from service during maintenance action. Namely, the crew is considered to be a train, so that part of the system is blocked for other trains. Another requirement, besides the safety, is to combine maintenance activities as much as possible. Moreover, it was decided to change the corrective maintenance into state dependent preventive maintenance. Doing so, the total amount of maintenance time will decrease and thus the inconvenience for the train operators decreases as well. Thus, the board of NS asked ProRail to divide the Dutch railway system into basic working zones and to develop an efficient maintenance schedule based on the working zone division. A working zone is part of the railways that is taken out of service when a maintenance crew is working on it.

Den Hertog *et al.* (2005) presents a method that has been used to divide the railway system in the Netherlands into working zones. The method takes the switches as basis. If one switch is out of operation the whole line segment behind it can no longer be used until the next switch. All maintenance activities in a certain working zone are combined and carried out during an 'out of service' for about five hours per night every four weeks. The working zones are indicated on maps. Every working zone has a unique code and these codes are on small signs attached to the rails. Van Zante-de Fokkert *et al.* (2007), the second part of the previous article, describes the method used for the construction of an optimal maintenance schedule. The most important restriction for constructing this schedule is that all maintenance activities take place within the working zones.

Switch maintenance is one of the most critical activities, since switches are often sources of failure and traffic disruption. Therefore, there should be a possibility to work on each switch every four weeks for about five hours net. In this way a four-week cyclic maintenance schedule has been developed, in which each working zone can be out of service for five and a half hours. The main lines (more than 300 working zones) are intensively used during the day; therefore taking out working zones of these lines in the daytime would cause severe disruptions for the railway traffic. Thus, the mentioned zones of the main lines are maintained at night (the passenger traffic is almost absent in the night and the cargo trains can be grouped: one set of trains per hour in each direction). To balance the maintenance contractor's workload between day and night, 10 percent of the zones of the main lines have been planned in the daytime. The maintenance schedule developed in Van Zante-de Fokkert *et al.* (2007) and used since 1999 in the Netherlands consists of three types of nights. Six nights are used only for planned maintenance, for eighteen nights no maintenance is scheduled (these nights can be used for track building and renewal projects) and finally, four nights can be used for both activities. This schedule is copied every four weeks. For large works the five and a half hours nightly time slots are not enough, therefore possessions can be incidentally scheduled in the weekends or during a couple of days/weeks in the low demand periods (holidays). This leads to the questions such as *What is the effect of such a weekend possession on the train operation? Will the train operation be completely canceled during such a weekend? Is it possible to restart the train operation after a weekend possession without any problems?*

In order to be able to answer these questions, we carried out a study together with a couple of researchers from NS. The results of this study are presented in Chapter 6. In the next section we give a short introduction into the subject of our research study with respect to the effect of the track possessions on the train operation.

1.3 Train operation during infrastructure maintenance

Currently, travelers are informed days/weeks beforehand if maintenance actions are planned for weekends or weekdays through announcements/leaflets on the stations, media and internet. They can get acquainted with traveling suggestions (*e.g.* travel via an other route, use slow trains instead of intercity trains), alternative traveling possibilities (*e.g.* travel by special buses) and extra traveling time (*e.g.* extra 30 minutes).

During track possessions the train operation is likely to be affected. At NS, for these days the timetable, the rolling stock and the crew schedules are adjusted and an *operational timetable*, an *operational rolling stock planning* and an *operational crew planning* is made. For a more detailed description of the planning process at NS, we refer to Maróti (2006) and Huisman *et al.* (2005). Thus, in the original (tactical) timetable trains are inserted, deleted or their departure and/or arrival times are changed. Furthermore, the type, number and order of the train units to be used for each train and the duties of the crew might be different during these maintenance periods than in the original (tactical) planning. Once the maintenance work is finished the train operation should be retaken as quick and smooth as possible. This means that the crew should retake their original duties, the trains should operate again according to their tactical timetable and with a composition prescribed in the tactical rolling stock planning. However, after a possession the rolling stock units may not finish their duties at the location where they were planned to or at the location where they have to start their duties next day. This is not a problem if two units of the same type get switched. In many cases, however, the number of units ending up after a possession at a certain station differs from the number of units that has to start their next day's duty there. To prevent expensive dead-heading trips, *i.e.* trips where empty units are moved from one location to the other one, it is attractive to consider this balancing issue already at the moment when the operational rolling stock schedule is made. This problem is called in Chapter 6 the Rolling Stock Rebalancing Problem (RSRP).

The input for RSRP consists of the operational timetable for a given planning period (*i.e.* possession), the available rolling stock types and the number of available units per type. Moreover, the *target inventories* are known, *i.e.* the number of units per type that are available at each station at the beginning of the planning period and the number of units per type that are needed at each station after the planning period. For example, assume that the planning period is the entire weekend (Saturday and Sunday). Then the target inventories for the rolling stock schedule are: the number of units that end at the stations on Friday evening and the number of units that start at the stations on Monday morning. Furthermore, we assume that there is an input rolling stock schedule, which meets all kind of requirements (these requirements are listed in Chapter 6) except that it may contain some *off-balances*. An off-balance is defined as a deviation from the target inventory level of a certain type of rolling stock at a certain station. For example, assume that at station A there are 2 units available before the infrastructure maintenance, while according to the input rolling stock schedule 3 units should depart from station A. Thus, station A has an off-balance. The primary goal in Chapter 6 of this thesis is to construct a new rolling stock schedule with as few off-balances as possible. As secondary objective, other criteria related to costs, service, and robustness may be optimized as well.

1.4 Thesis objectives

This thesis can be divided into two parts. In Part I we are dealing with the problem of finding optimal time intervals for carrying out routine maintenance works and large projects in such a way that the track possession costs and maintenance costs are minimized. In Part II of this thesis we focus on rescheduling of the rolling stock in the passenger railways due to changing circumstances and more precisely on the Rolling Stock Rebalancing Problem (RSRP). This problem arises within a passenger railway operator when the rolling stock has to be re-scheduled due to planned maintenance of the railway infrastructure, unplanned disruptions of the railway system, *etc.*

The main objectives of this thesis are formulated as follows:

1. Review the existing literature on maintenance planning in relation with production.
In this thesis, Chapter 2 presents an overview of mathematical models that consider the relation between maintenance and production. Apart from giving a general overview of the models we also discuss briefly some business sectors in which the interactions between maintenance and production have been studied.

2. Identify some tactical and operational railway infrastructure maintenance planning problems and develop operations research models for providing decision support. Investigate the effect of planning railway infrastructure maintenance on the train operation and identify rolling stock planning problems that occur during planned infrastructure mainte-

nance.
One of the sectors that we examine in Chapter 2 is the railway sector. In Chapter 3 we investigate in more detail the literature on maintenance planning on the railway infrastructure. During both literature studies we learned that the railway infrastructure maintenance is very difficult to plan and involves high costs. Therefore there is a need for developing operations research tools, which help the maintenance planners to come up with optimal maintenance plans. This led to the models and solution methods presented in Chapter 4 and Chapter 5. Moreover, in Chapter 6 we investigated the rolling stock rebalancing problem that occurs when the rolling stock has to be re-scheduled due to changes in the timetable, such as planned infrastructure maintenance.

3. Analyze the considered models, investigate their computational complexity, propose solution methods and test the solutions of the models.
In Chapters 4, 5 and 6 of this thesis we formulate several models as mixed integer linear programs. For the implementation of these models we make use of GAMS (General Algebraic Modeling System, GAMS homepage (2009)) as well as CPlex software (see ILOG homepage (2009)). CPlex is used both as an Interactive Base System operating with command lines and as a Callable Library of routines which could be called from the programs developed by us. Furthermore, some of the problems are solved using the genetic and memetic algorithms (Chapter 5), including advanced local search algorithms as tabu search, simulated annealing and steepest hill climbing. Besides this, we propose different type of heuristics and study their behavior on randomly generated instances (Chapters 4 and 5) and on real life instances (Chapter 6). The models and solution methods from this thesis are developed for solving planning problems in the railways, but we believe that many of these solution approaches, especially from Chapters 4 and 5, can be applied in other business sectors as well.

1.5 Outline of the thesis

This thesis consists of five central chapters, this introduction (Chapter 1) and a final chapter (Chapter 7), where the general conclusions are summarized. The five central chapters are based on four papers on maintenance optimization and one paper on operational rolling stock circulation. Although chapters 2, 3, 4 and 5 are interrelated, as they all deal with maintenance optimization, they are the result of independent research projects. Thus, each chapter has its own introduction to the topic of its scope, and presents the literature review relevant to it. As a results, it is possible to find some overlapping material in chapters 3, 4 and 5. The references for the chapters are listed in the bibliography at the end of the thesis.

Chapter 2 consists of an overview of mathematical models that consider the relation between maintenance and production. The relation exists in several ways. First of all, when planning maintenance one needs to take production into account. Secondly, maintenance can also be seen as a production process, which needs to be planned and finally one can develop integrated models for maintenance and production. Next to giving a general overview of models we also discuss some sectors (railway, road, airline, electric power) in which the interactions between maintenance and production have been studied. This chapter is based on Budai, Dekker and Nicolai (2008).

In Chapter 3, that is based on Budai and Dekker (2002), we give a literature review on the planning of maintenance of railway infrastructure and especially on the effect of maintenance on railway capacity. The review is based on a general structuring of maintenance planning in eight phases. This research has been carried out in the framework of the European "IMPROVERAIL" project (see Improverail (2002)), an international research consortium "*supporting the establishment of railway infrastructure management in accordance with Directive 91/440, by developing the necessary tools for the modelling of railway infrastructure and access management*".

From the literature studies we could conclude that railway infrastructure maintenance is very difficult to plan since there are many constraints to be considered. Firstly, carrying out maintenance on the rail infrastructure usually involves many disturbances for the travelers (*e.g.* delays, canceled trains), and vice versa, the train operation restricts the length and the frequency of the infrastructure possession. Moreover, there are high costs involved. Nevertheless, Chapter 4 of this thesis deals with the Preventive Maintenance Scheduling Problem (PMSP) for the railway infrastructure, where routine maintenance activities and unique projects have to be scheduled in a certain time period such that the sum of the maintenance costs and the possession costs is minimized. The possession costs are mainly determined by the possession time. First, we provide a math programming formulation for the PMSP and we prove that PMSP is NP-hard. After that some heuristics have been developed to solve this problem fast. This chapter is based on Budai, Huisman and Dekker (2006).

Chapter 5 extends the research done in the previous chapter by using the genetic and memetic algorithms to solve the PMSP. The performance of these algorithms are compared with each other and with the performance of the methods developed in the previous chapter. This chapter is based on Budai, Dekker and Kaymak (2009b).

Chapter 6 deals with the Rolling Stock Rebalancing Problem (RSRP), which arises at various stages of the planning process of a passenger railway operator, namely from the short-term planning phase (*i.e.* planning some days or weeks ahead) till the real-time operations. Due to changes in the timetable (*e.g.* planned maintenance or unplanned disruptions) the previously created rolling stock schedules have to be adjusted for a certain time period. A two-phase heuristic approach has have been developed to solve the RSRP.

The performances of this heuristic is compared with the performance of an iterative heuristic developed in Maróti (2006) and with the performance of the exact solution method of Fioole *et al.* (2006) used at NS. For the purpose of making the comparison of the two heuristics more clear, we describe in Chapter 6 the iterative heuristic too. This chapter is based on Budai, Maróti, Dekker, Huisman and Kroon (2009a).

Finally, in Chapter 7 we draw some conclusions and indicate possible directions for further research on the topics considered in this thesis.

Chapter 2

Maintenance & Production: A Review of Planning Models

Abstract

In this chapter we give an overview of the relation between planning of maintenance and production. First we consider production planning and scheduling models where failures and maintenance aspects are taken into account. Next we discuss the planning of maintenance activities, where we consider both preventive as well as corrective maintenance. Thirdly, we consider the planning of maintenance activities at such moments in time where the items to be maintained are not or less needed for production. This type of maintenance is also called opportunity maintenance. Apart from describing the main ideas, approaches and results we also provide a number of applications.

This chapter is based on Budai, Dekker and Nicolai (2008).

2.1 Introduction

Maintenance is the set of activities carried out to keep a system into a condition where it can perform its function (British Standard (1984)). Quite often these systems are production systems where the output are products and/or services. Some maintenance can be done during production and some can be done during regular production stops in evenings, weekends and on holidays. However, in many cases production units need to be shut down for maintenance. This may lead to tensions between the production and maintenance department of a company. On one hand the production department needs maintenance for the long-term well-being of their equipment, on the other hand it leads to shutting down the operations and loss of production. It will be clear that both can benefit from decision support based on mathematical models.

In this chapter we give an overview of mathematical models that consider the relation between maintenance and production. The relation exists in several ways. First of all, when planning maintenance one needs to take production into account. Secondly, maintenance can also be seen as a production process which needs to be planned and finally one can develop integrated models for maintenance and production. Apart from giving a general overview of models we will also discuss some sectors in which the interactions between maintenance and production have been studied.

Many review articles have been written on maintenance, *e.g.* Cho and Parlar (1991), but to our knowledge only one on the combination between maintenance and production, Ben-Daya and Rahim (2001). Our review differs from that one in several aspects. First of all, we also consider models which take production restrictions into account, rather than integrated models. Secondly we discuss some specific sectors. Finally, we discuss the more recent articles since that review.

The articles that we refer to in this study were found using Google Scholar, Scirus and Scopus as search engines and ScienceDirect, JStor, MathSciNet and Emerald as online databases. We primarily searched on keywords, abstracts and titles, but we also searched within the articles for relevant references. Overview articles on maintenance (*e.g.* Cho and Parlar (1991), Dekker (1996), and Wang (2002)) and the review of Ben-Daya and Rahim (2001) on production, maintenance and quality models have been very useful sources for detecting important articles in different fields within production planning and/or maintenance scheduling. Moreover, we have applied a citation search (looking both backwards in time and forwards in time for citations) to all articles found. This citation search is an indirect search method, whereas the above methods are direct methods. The advantage of this method is that one can easily distinguish clusters of related articles. Note that studies published in books or proceedings that are not electronically available, are likely to have escaped.

The following terms and/or a combination of them were used for searching the literature: maintenance, maintenance optimization, production, production planning, downtime cost, airlines, railways, road, process industry, opportunity, manpower.

Classification scheme

Maintenance is related to production in several ways. First of all, maintenance is intended to allow production, yet to execute maintenance production often has to be stopped. Therefore, this negative effect has to be considered in maintenance planning and optimization. It comes specifically forward in the costing of downtime and in opportunity maintenance. All articles taking the effect of production on maintenance explicitly into account fall into this category.

Secondly, maintenance can also be seen as a production process which needs to be planned. Planning in this respect implies determining appropriate levels of capacity (*e.g.* manpower) concerning the demand.

Thirdly, we are concerned with production planning in which one needs to take maintenance jobs into account. The point is that the maintenance jobs take production capacity away and hence they need to be planned together with production. Maintenance has to be done either because of a failure or because the quality of the produced items is not good enough. In this third category we also consider the integrated planning of production and maintenance.

The relation between maintenance and production is also determined by the business sector. We consider the following sectors: railway, road, airlines and electrical power generation maintenance.

The outline of this chapter is now as follows. In Section 2.2 we present an overview of the main elements of maintenance planning, as these are essential to understand the rest of this chapter. Following our classification scheme, in Section 2.3 we review articles in which maintenance is modelled explicitly and where the needs of production are taken into account. Since these needs differ between business sectors, we discuss in Section 2.4 the relation between production and maintenance for some specific business sectors. In Section 2.5 we consider the second category in our classification scheme: maintenance as a production process which needs to be planned. In Section 2.6 we are concerned with production planning in which one needs to take maintenance jobs into account (integrated production and maintenance planning). Trends and open research areas are discussed in Section 2.7 and finally, in Section 2.8 conclusions are drawn.

2.2 Maintenance planning and optimization: a recap

In maintenance several important decisions have to be made. We distinguish between (i) the long term (strategic) planning, (ii) medium term (tactical) planning and finally (iii) short term scheduling.

Major strategic decisions concerning maintenance are made in the design process of systems. What type of maintenance is appropriate and when should it be done? This is laid down in the so-called maintenance concept. Many optimization models address this problem and the relation with production is implicitly covered by them.

Another important strategic problem is the organization of the maintenance department. Is maintenance done by production personnel or by a specific maintenance personnel? Secondly, questions such as “Where is it located?” “Are specific types of work outsourced?”, et cetera, should be answered. Although they are important topics, they are more the concern of industrial organization than the topic of mathematical models.

Further important strategic issues concern how a system can be maintained, whether specific expertise or equipment needed, whether one can easily reach the subsystems, what information is available and what elements can be easily replaced. These are typical maintainability aspects, but they have little to do with production.

In the tactical phase, usually between a month and year, one makes a plan for the major maintenance/upgrade of major units and this has to be done in cooperation with the production department. Accordingly, specific decision support is needed in this respect.

Another tactical problem concerns the capacity of the maintenance crew. Is there enough manpower to carry out the preventive maintenance program? These questions can be addressed by use of models as will be indicated later on.

In the short term scheduling phase one determines the moment and order of execution, given an amount of outstanding corrective or preventive work. This is typically the domain of work-scheduling, where extensive model-based support can be given.

We will next consider another important aspect in maintenance, which is the type of maintenance. A typical distinction is made between corrective and preventive maintenance work. The first is carried out after a failure, which is defined as the event by which a system stops functioning in a prescribed way. Preventive work however, is carried out to prevent failures. Although this distinction is often made, we like to remark that the difference is not that clear as it may seem. This is due to the definition of failure. An item may be in a bad state, while still functioning and one may or may not consider this as a failure. Anyhow, an important distinction between the two is that corrective maintenance is usually not plannable, but preventive maintenance typically is.

The execution of maintenance can also be triggered by condition measurements and then we speak over condition-based maintenance. This has often been advocated as more effective and efficient than time-based preventive maintenance. Yet it is very hard to predict failures well in advance, and hence condition-based maintenance is often unplannable. Instead of time based maintenance one can also base the preventive maintenance on utilization (run hours, mileage) as being more appropriate indicators of wear out. Finally, one may also have inspections which can be done by sight or instruments and often do not affect operation. They do not improve the state of a system however, but only the information about it. This can be important in case machines may start producing items of a bad quality. There are inspection-quality problems where inspection optimization is connected to quality control.

Another distinction is about the amount of work. Often there are small works, grouped into maintenance packages. They may start with inspection, cleaning and next some improvement actions, like lubricating and or replacing some parts. These are typically part of the preventive maintenance program attached to a system and have to be done on a repetitive basis (month, quarter, year or two-years). Next, one has replacements of parts/subsystems and overhauls or refurbishments, where a system is substantially improved. The latter are planned long ahead and carried out as projects with individual budgets.

In classical maintenance optimization the objective is to find a trade-off between preventive and corrective maintenance. The typical motivation is that preventive maintenance is cheaper than corrective. Usually, maintenance costs are due to manhours, materials and indirect costs. The difference between corrective and preventive maintenance costs is especially in the indirect costs. They consist of loss of production, environmental damages or safety consequences. Costing these consequences can be a difficult problem and it is tackled in Section 2.3.1. It will also be clear that preventive maintenance should be performed when production is least affected. One way of achieving it is by carrying out maintenance actions at opportunities. This has given rise to a specific class of models described in Section 2.3.2.

2.3 When to do maintenance in relation with production

In this section we discuss articles in which maintenance planning is modelled explicitly and the needs of production are taken into account. The latter however, is not usually modelled as such, but it is taken into account in the form of constraints or requirements. Alternatively, the effect of maintenance on varying production scenarios may be considered. Following this reasoning we arrive at three streams of research. A first stream assesses the costs of downtime, which is important in the planning of maintenance. The second stream deals with studies where one tries to schedule maintenance work at those moments that units are not needed for production (opportunities) and in the last stream articles are considered which schedule maintenance in line with production. Each stream is discussed in a separate section.

2.3.1 Costing of downtime

Assessing the costs of downtime is an important step in the determination of costs of preventive and corrective maintenance. Although exact values are not necessary, as most optimization results show, it is important to assess these values with a reasonable accuracy. It is easier to determine downtime costs in case of preventive maintenance than in case of corrective maintenance as failures may have many unforeseen consequences. Yet even in case of preventive maintenance the assessment can be difficult, *e.g.* in case of highway shutdowns or railway stoppage.

Another problem to be tackled is the system-unit relation. A system can be a complex configuration of different units, which may imply that downtime of one unit does not necessarily halt the full system. Accordingly, an assessment of the consequences of unit downtime on system performance has to be made. This is especially a problem in case of k-out-of-n systems or even in more general configurations.

Several articles deal with this issue. Some give an overall model, others describe a detailed case. Geraerds (1985) gives an outline of a general structuring to determine downtime costs. In Dekker and van Rijn (1996) a downtime model is described for k-out-of-n systems used on the oil production platforms. Edwards *et al.* (2002) give a detailed model for the costs of equipment downtime in open-pit mining. They use regression models based on historical data. Furthermore, Knights *et al.* (2005) present a model to assist maintenance managers in evaluating the economic benefits of maintenance improvement projects.

2.3.2 Opportunity maintenance

Opportunity maintenance is the maintenance that is carried out at an opportune moment, *i.e.* moments at which the units to be maintained are less needed for their function than normally. We speak of opportunities if these events occur occasionally and if they are difficult to predict in advance. There can be several reasons for maintenance opportunities:

- Failure and hence repairs of other units/components.

  The failure of one component often is an opportunity to preventively maintain other components. Especially if the failure causes the breakdown of the production system it is favorable to perform preventive maintenance on other components as well. After all, little or no production is lost above the one resulting from the original failure. An example is given in Van der Duyn Schouten *et al.* (1998), who consider the replacement of traffic lights at an intersection.

- Other interruptions of production.

  Production processes are not only interrupted by failures or repairs. Several outside events may create an opportunity as well. This can be market interruptions, or other work by which production needs to be stopped (*e.g.* replacing catalysts *etc.*).

According to the foregoing discussion, the articles dealing with opportunity maintenance are divided into two categories. In the articles from the first category it is assumed that upon a failure preventive maintenance is carried out on other components of the system as well. In the articles from the second category the opportunities are modelled as an outside event at which one may do maintenance.

Bäckert and Rippin (1985) consider the first type of opportunistic maintenance for plants subject to breakdowns. In this article three methods are proposed to solve the problem. In the first two cases the problem is formulated as a stochastic decision tree and solved using a modified branch and bound procedure. In the third case the problem is formulated as a Markov decision process. The planning period is discretized, resulting in a finite state space to which a dynamic programming procedure can be applied.

In Wijnmalen and Hontelez (1997) a multi-component system is considered, where failures of one component may create an opportunity for the other components of the system. The opportunity process is approximated by an independent process with the same mean rate. In this way they circumvent the problem of dimensionality, which appears in the study of Bäckert and Rippin (1985). The last article from the first category that we mention here is Tan and Kramer (1997). The authors propose a general framework for preventive maintenance optimization in chemical process operations, where Monte Carlo simulation is combined with a genetic algorithm.

There are several articles considering the second stream. In Dekker and Dijkstra (1992) and Dekker and Smeitink (1991) it is assumed that the opportunity-generating process is completely independent of the failure process and is modelled as a renewal process. Dekker and Smeitink (1994) consider multi-component maintenance at opportunities of restricted duration and determine priorities of what preventive maintenance to do at an opportunity.

In Dekker and van Rijn (1996) a decision-support system (PROMPT) for opportunity-based preventive maintenance is discussed. PROMPT was developed to take care of the random occurrence of opportunities of restricted duration. Here, opportunities are not only failures of other components, but also preventive maintenance on (essential) components. Many of the techniques developed in the articles of Dekker and Smeitink (1991), Dekker and Dijkstra (1992) and Dekker and Smeitink (1994) are implemented in the decision-support system. In PROMPT preventive maintenance is split up into packages. For each package an optimum policy is determined that indicates when it should be carried out. Furthermore, a priority measure is determined that decides which maintenance package should be executed at a given opportunity.

In Dekker *et al.* (1998b) the maintenance of light-standards is studied. A light-standard consists of n independent and identical lamps screwed on a lamp assembly. To guarantee a minimum luminance, the lamps are replaced if the number of failed lamps reaches a prespecified number m. In order to replace the lamps the assembly has to be lowered. As a consequence, each failure is an opportunity to combine corrective and preventive maintenance. Several opportunistic age-based variants of the m-failure group replacement policy are considered. Simulation optimization is used to determine the optimal opportunistic age threshold.

Dagpunar (1996) introduces a maintenance model where replacement of a component within a system is possible when some other part of the system fails, at a cost of c_2. The opportunity process is Poisson. A component is replaced at an opportunity if its age exceeds a specified control limit t. Upon failure a component is replaced at cost c_4 if its age exceeds a specified control limit x, otherwise it is minimally repaired at cost c_1. In case of a minimal repair the age and failure rate of the component after the repair is as it was immediately before failure. There is also a possibility of a preventive or "interrupt" replacement at cost c_3 if the component is still functioning at a specified age T. A procedure to optimize the control limits t and T is given in Dekker and Plasmeijer (2001).

2.3.3 Maintenance scheduling in line with production

In this section we consider models where the effect of production on maintenance is explicitly taken into account. These models only address maintenance decisions, so they do not give advice on how to plan production.

The models developed in the articles in this category show that a good maintenance plan, one that is integrated with the production plan, can result in considerable cost savings. This integration with production is crucial because production and maintenance have a direct relationship. Any breakdown in machine operation results in disruption of production and leads to additional costs due to downtime, loss of production, decrease in productivity and quality, and inefficient use of personnel, equipment and facilities. Below we review articles following this stream of research in chronological order.

Dedopoulos and Shah (1995) consider the problem of determining the optimal preventive maintenance policy parameters for individual items of equipment in multipurpose plants. In order to formulate maintenance policies, the benefits of maintenance, in the form of reduced failure rates, must be weighed against the costs. The approach in this study first attempts to estimate the effect of the failure rate of a piece of equipment on the overall performance/profitability of the plant. An integrated production and maintenance planning problem is also solved to determine the effects of preventive maintenance on production. Finally, the results of these two procedures are then utilized in a final optimization problem that uses the relationship between profitability and failure rate as well as the costs of different maintenance policies to select the appropriate maintenance policy.

Vatn *et al.* (1996) present an approach for identifying the optimal maintenance schedule for the components of a production system. Safety, health and environment objectives, maintenance costs and costs of lost production are all taken into consideration, and maintenance is thus optimized with respect to multiple objectives. The approach is flexible as it can be carried out at various levels of detail, *e.g.* adopted to available resources and to the management's willingness to give detailed priorities with respect to objectives on safety versus production loss.

Frost and Dechter (1998) define the scheduling of preventive maintenance of power generating units within a power plant as constraint satisfaction problems. The general purpose of determining a maintenance schedule is to determine the duration and sequence of outages of power generating units over a given time period, while minimizing operating and maintenance costs over the planning period.

Vaurio (1999) develops unavailability and cost rate functions for components whose failures can occur randomly. Failures can only be detected by periodic testing or inspections. If a failure occurs between consecutive inspections, the unit remains failed until the next inspection. Components are renewed by preventive maintenance periodically, or by repair or replacement after a failure, whichever occurs first. The model takes into account finite repair and maintenance durations as well as costs due to testing, repair, maintenance and lost production or accidents. For normally operating units the time-related penalty is loss of production. For standby safety equipment it is the expected cost of an accident that can happen when the component is down due to a dormant failure, repair or maintenance. The objective is to minimize the total cost rate with respect to the inspection and the replacement interval. General conditions and techniques are developed for solving optimal test and maintenance intervals, with and without constraints on the production loss or accident rate. Insights are gained into how the optimal intervals depend on various cost parameters and reliability characteristics.

Van Dijkhuizen (2000) studies the problem of clustering preventive maintenance jobs in a multiple set-up multi-component production system. As far as we know, this is the first attempt to model a maintenance problem with a hierarchical (tree-like) set-up structure. Different set-up activities have to be done at different levels in the production system before maintenance can be done. Each component is maintained preventively at an integer multiple of a certain basis interval, which is the same for all components, and corrective maintenance is carried out in between whenever necessary. So, every component has its own maintenance frequency. The frequencies are based on the optimal maintenance planning for single components. Obviously, set-up activities may be combined when several components are maintained at the same time. The problem is to find the maintenance frequencies that minimize the average cost per unit of time.

Cassady *et al.* (2001) introduce the concept of selective maintenance. Often production systems are required to perform a sequence of operations with finite breaks between each operation. The authors establish a mathematical programming framework for assisting decision-makers in determining the optimal subset of maintenance activities to perform prior to beginning the next operation. This decision making process is referred to as selective maintenance.

The article of Haghani and Shafahi (2002) deals with the problem of scheduling bus maintenance activities. A mathematical programming approach to the problem is proposed. This approach takes as input a given daily operating schedule for all buses assigned to a depot along with available maintenance resources. Then a daily inspection and maintenance schedule is designed for the buses that require inspection so as to minimize the interruptions in the daily bus operating schedule, and maximize the reliability of the system and efficiently utilize the maintenance facilities.

Charles *et al.* (2003) examine the interaction effects of maintenance policies on batch plant scheduling in a semiconductor wafer fabrication facility. The purpose of the work is the improvement of the quality of maintenance department activities by the implementation of optimized preventive maintenance strategies. The production of semiconductor devices is carried out in a wafer lab. In this production environment equipment breakdown or procedure drifting usually induces unscheduled production interruptions.

Cheung *et al.* (2004) consider a plant with several units of different types. There are several shutdown periods for maintenance. The problem is to allocate units to these periods in such a way that production is least affected. Maintenance is not modelled in detail, but incorporated through frequency or period restrictions.

2.4 Specific business sectors

The purpose here is to illustrate the interdependence between maintenance and production for some specific sectors in more detail. Moreover, it shows what ideas were employed in which sector and the difference between them. Although many sectors could be distinguished we take those where maintenance plays an important role. Not surprisingly, these are all capital intensive sectors with high maintenance expenditure. Below we discuss railway, road, airline and electric power generation maintenance.

2.4.1 Railway maintenance

Since rail is an important transportation mode, proper maintenance of the existing lines, repairs and replacements carried out in time are all important to ensure efficient operation. Moreover, since some failures might have a strong impact on the safety of the passengers, it is important to prevent these failures by carrying out in time and according to some predefined schedules preventive maintenance works. The preventive maintenance works are the small routine works and/or projects. The routine (spot) maintenance activities, that consist of inspections and small repairs (see Esveld (2001)), do not take much time to be performed and are done regularly, with frequencies varying between monthly and once a year. The projects include renewal works and they are carried out once or twice every few years.

In the literature there are a couple of articles that provide useful methods for finding optimal track possession intervals for carrying out preventive maintenance works, *i.e.* time periods when a track is required for maintenance, therefore it will be blocked for the operation. In production planning terms track possession means downtime required for maintenance. The main question is when to carry out maintenance such that the inconvenience for the train operators, the disruption to and from the scheduled trains, the infrastructure possession time for maintenance are minimized and the maintenance cost is the lowest possible. For a more detailed overview of techniques used in planning railway infrastructure maintenance we refer to Budai and Dekker (2002) and Improverail (2002). In some articles (Higgins (1998), Cheung *et al.* (1999) and Budai *et al.* (2006)) the track possession is modelled in between operations. This can be done for occasionally used tracks, which is the case in Australia and some European countries. If tracks are used frequently, one has to perform maintenance during nights, when the train traffic is almost absent, or during weekends (with possible interruption of the train services), when there are fewer disturbances for the passengers. In the first case one can either make a cyclic static schedule, which is done by Den Hertog *et al.* (2005) and Van Zante-de Fokkert *et al.* (2007) for the Dutch situation, or a dynamic schedule with a rolling horizon, which is presented in Cheung *et al.* (1999). The latter schedule has to be made regularly.

Some other articles deal with grouping railway maintenance activities to reduce costs, downtime and inconvenience for the travellers and operators. Here we mention the study of Budai *et al.* (2006) in which the preventive maintenance scheduling problem is introduced. This problem arises in other public/private sectors as well, since preventive maintenance of other technical systems (machine, road, airplanes, *etc.*) also contains small routine works and large projects.

2.4.2 Road maintenance

Road maintenance and railway maintenance have many common characteristics. Failures are often indirect, in the sense that norms are surpassed, but there may not be any consequences. The production function is indirect, but that does not mean that it is not felt by many. Governments may define a cost penalty due to one hour waiting per vehicle because of congestion caused by road maintenance. Similar as in railway maintenance, one sees that work is shifted to nights or a lot of work is combined into a large project on which the public is informed long before it is started. The night work causes high logistics costs for maintenance, yet it is useful for small repairs or patches.

Other similarities with railroads are the large number of identical parts (a road is typically split up in lanes of 100 meters about which information is stored). Vans with complex road analyzing equipment are used to assess the road quality. For railways special trains with complex measuring equipment are used. Videos are used in both cases. Next, both roads and rails have multiple failure modes. Furthermore, the assets to be maintained are spread out geographically, which results in high logistics costs for maintenance. This is also true for airline and truck maintenance. Both road and rail need much maintenance and large budgets need to be allocated for both.

Although several articles have been written on road maintenance, few take the production or user consequences into account. We like to mention Dekker *et al.* (1998a) which compares two concepts to do road maintenance. One with small work during nights and the other where large road segments (some 4 km) are overhauled in one stretch. In the latter case the traffic is diverted to other lanes or sides of the road. It is shown that the latter is both advantageous for the traffic as well as cheaper, provided that the road is not congested. Another interesting contribution is from Rose and Bennett (1992) who provide a model to locate and decide on the size (or capacity) of road maintenance depots for corrective maintenance.

2.4.3 Airline maintenance

Maintenance costs are a substantial factor of airline costs. Estimates are that 20% of the costs is due to maintenance. Maintenance is crucial because of safety reasons and because of high downtime costs. Besides a crash, the worst event for an airline is an aircraft on ground because of failures. Accordingly, a lot of technology has been developed to facilitate maintenance. Here we mention in-flight diagnosis, such that quick actions can be taken on ground and a very high level of modularity, such that failed components can easily be replaced. Moreover, in an aircraft there is a high level of time-based preventive maintenance rather than condition-based maintenance. A plane has to undergo several checks, ranging from an A check taking about an hour after each flight, to a monthly B check, a yearly C check and a five-yearly D check, where it is completely overhauled and which can last for a month. The presence of the monthly check implies that planes can not always fly the same route, but need to be rotated on a regular basis. It also implies that airlines need multiple units of a type in order to provide a consistent service.

Several studies have addressed the issue of fleet allocation and maintenance scheduling. In the fleet allocation one decides which planes fly which route and at which time. One would preferably make an allocation which remains fixed for a whole year, but due to the regular maintenance checks this is not possible. Gopalan and Talluri (1998) give an overview of mathematical models on this problem. El Moudani and Mora-Camino (2000) present a method to do both flight assignment as well as maintenance scheduling of planes. It uses dynamic programming and heuristics. A case of a charter airline is considered. Sriram and Haghani (2003) also consider the same problem. They solve it in two phases. Finally, Feo and Bard (1989) consider the problem of maintenance base planning in relation to an airline fleet rotation, while Cohn and Barnhart (2003) consider the relation between crew scheduling and key maintenance routing decisions.

In an other line of research, Dijkstra *et al.* (1994) develop a model to assess maintenance manpower scheduling and requirements in order to perform inspection checks (the A type) between flight turnarounds. It appears that their workload is quite peaked, because of many flights arriving more or less at the same time (so-called banks) in order to allow fast passenger transfers. The same problem is also tackled later by Yan *et al.* (2004).

As the last article in this category we like to mention Cobb (1995), who presents a simulation model to evaluate current maintenance system performance or the positive effect of ad hoc operating decisions on maintenance turn times (*i.e.* the time that maintenance takes to carry out a check or to do a repair).

2.4.4 Electric power system maintenance

Kralj and Petrovic (1988) present an overview article on optimal maintenance of thermal generating units in power systems. They primarily focused on articles published in IEEE Transactions on Power Apparatus and Systems. Here we will briefly discuss the typical problems of the maintenance of power systems and review two articles dealing with these problems.

First of all, note that maintenance of power systems is costly, because it is impossible to store generated electrical energy. Moreover, the continuity of supply is very important for its customers. A second problem of scheduling the maintenance of power systems is that joint maintenance of units is often impossible or very expensive, since that would affect production too much.

Frost and Dechter (1998) consider the problem of scheduling preventive maintenance of power generating units within a power plant. The purpose of the maintenance scheduling is to determine the duration and sequence of outages of power generating units over a given time period, while minimizing operating and maintenance costs, that are subjects to various constraints. A subset of the constraints contains the pairs of components that cannot be maintained simultaneously. In this article the maintenance problems are cast as constraint satisfaction problems (CSP). The optimal solution is found by solving a series of CSPs with successively tighter cost-bound constraints.

Langdon and Treleaven (1997) study the problem of scheduling maintenance for electrical power transmission networks. Grouping maintenance in the network may prevent the use of a cheap electricity generator, requiring a more expensive generator to be run in its place. Therefore, some parts of the network should not be maintained simultaneously. These exclusions are modelled by adding restrictions to the MIP formulation of the problem.

2.5 Production planning of maintenance

Maintenance can also be regarded as a production process which needs to be planned. Planning in this respect implies determining appropriate levels of capacity concerning the demand. It will be clear that this activity can only be carried out for plannable maintenance, *e.g.* overhauls or refurbishment and that it is only needed when there are capacity restrictions, *e.g.* in shipyard.

Articles in this category are Dijkstra *et al.* (1994) and Yan *et al.* (2004), who both consider manpower determination and allocation problems in case of a fluctuating workload for aircraft maintenance. Shenoy and Bhadury (1993) use the MRP approach to develop a maintenance manpower plan. Bengü (1994) discusses the organization of maintenance centers that are specialized to carry out particular types of maintenance jobs in the telecommunication sector. Al-Zubaidi and Christer (1997) consider the problem of manpower planning for hospital building maintenance.

Another typical production planning problem is with respect to layout planning. A case study for a maintenance tool room is described in Rosa and Feiring (1995). The study by Rose and Bennett (1992), which was discussed in Section 2.4.2, also falls into this category.

2.6 Integrated production and maintenance planning

In recent years there has been considerable interest in models attempting to integrate production, quality and maintenance (Ben-Daya and Rahim (2001)). Whereas in the past these aspects have been treated as separate problems, nowadays models take the mutual interdependencies into account. Production planning typically concerns determining lot sizes and evaluating capacity needs in case of fluctuating demand. Both the optimal lot size and the capacity needs are influenced by failures. On the other hand, maintenance prevents breakdowns and improves quality. Accordingly, they should be planned in an integrated way (see *e.g.* Nahmias (2005)).

We subdivide the class of integrated production and maintenance planning models in four categories: high-level models considering conceptual and process design problems (Section 2.6.1); the economic manufacturing quantity model, which was originally posed as a simple inventory problem, but has been (successfully) extended to deal with quality and failure aspects (Section 2.6.2); models of production systems with buffer capacities, which by definition are suitable to deal with breakdowns (Section 2.6.3); finally, production and maintenance rate optimization models, which aim to find the production and preventive/corrective maintenance rates of machines so as to minimize the total cost of inventory, production and maintenance (Section 2.6.4). In Section 2.6.5 we discuss articles which do not fit in any of the above categories.

2.6.1 Conceptual and design models

In a number of articles conceptual models are developed that integrate the preventive and corrective aspects of the maintenance planning, with aspects of the production system such as quality, service level, priority and capacity activities. For instance, Finch and Gilbert (1986) present an integrated conceptual framework for maintenance and production in which they focus especially on manpower issues in corrective and preventive work. Weinstein and Chung (1999) test the hypothesis that integrating the maintenance policy with the aggregate production planning will significantly influence total cost reduction. It appears that this is the case in the experimental setting investigated in this study. Lee (2005) considers production inventory planning, where high level decisions on maintenance (viz. their effects) are made.

Another group of articles deal with integrating process design, production and maintenance planning. Already at the design stage decisions on the process system and on initial reliability of the equipments are made. Pistikopoulos *et al.* (2000) describe an optimization framework for general multipurpose process models, which determine both the optimal design as well as the production and maintenance plans simultaneously. In this framework, the basic process and system reliability/maintainability characteristics are determined in the design phase with the selection of system structure, components, *etc.* The remaining characteristics are determined in the operation phase with the selection of appropriate operating and maintenance policies. Therefore, the optimization of process system effectiveness depends on the simultaneous identification of optimal design, operation and maintenance policies having properly accounted for their interactions. In Goel *et al.* (2003) a reliability allocation model is coupled with the existing design, production, and maintenance optimization framework. The aim is to identify the optimal size and initial reliability for each unit of equipment at the design stage. They balance the additional design and maintenance costs with the benefits obtained due to increased process availability.

2.6.2 EMQ problems

In the classical economic manufacturing quantity (EMQ) model items are produced at a constant rate p and the demand rate for the items is equal to $d < p$. The aim of the model is to find the production uptime that minimizes the sum of the inventory holding cost and the average, fixed, ordering cost. This model is an extension of the well-known economic order quantity (EOQ) model; the difference being that in the EOQ model orders are placed when there is no inventory. Note that the EMQ model is also referred to as economic production quantity (EPQ) model.

In the extensive literature on production and inventory problems, it is often assumed that the production process does not fail, that it is not interrupted and that it only produces items of acceptable quality. Unfortunately, in practice this is not always the case. A production process can be interrupted due to a machine breakdown or because the quality of the produced items is not acceptable anymore. The EMQ model has been extended to deal with these aspects and we thus divide the literature on EMQ models into two categories. First, we consider EMQ problems that take into account the quality aspects of the items produced. The second category of EMQ models analyzes the effects of (stochastic machine) breakdowns on the lot sizing decision.

2.6.2.1 EMQ problems with quality aspects

One of the reasons why a production process is interrupted is the (lack of) quality of the items produced. Obviously, items of inferior quality can only be sold at a lower revenue or cannot be sold at all. Thus, the production of these items results in a loss (or a lower profit) for the firm. This type of interruption is usually modelled as follows. It is assumed that at the start of the production cycle the production is in an "in-control" state, producing items of acceptable quality. After some time the production process may then shift to an "out-of-control" state. In this state a certain percentage of the items produced are defective or of sub-standard quality. The elapsed time for the process to be in the in-control state, before the shift occurs, is a random variable. Once a shift to the out-of-control state has occurred, it is assumed that the production process stays in that state unless it is discovered by (a periodic) inspection of the process, followed by corrective maintenance.

One of the earliest works that consider the problem of finding the optimal lot size and optimal inspection schedule is the article of Lee and Rosenblatt (1987). They show that the derived optimal lot size is smaller than the classical EMQ if the time for the process to be in the in-control state follows an exponential distribution. Lee and Rosenblatt (1989) have extended this work by assuming that the cost of restoration is a function of the elapsed time since a shift from an in-control to an out-of-control state of the production process has occurred. In addition, the possibility of incurring shortages in the model is allowed.

Many attempts have been made to extend these two models. For instance, Tseng (1996) assumes that the process lifetime is arbitrarily distributed with an increasing failure rate. Furthermore, two maintenance actions are considered. The first is a perfect maintenance action, which restores the system to an as-good-as new condition if the process is in the in-control state. If however, the production process is in out-of-control state, it is restored to the in-control state at a given restoration cost. Secondly, maintenance is always done at the end of a production cycle to ensure that the process is perfect at the beginning of each production cycle.

Wang and Sheu (2003) assume that the periodic inspections are imperfect. Two types of inspection errors are considered, namely (I) the process is declared out-of-control when it is in-control and (II) the process is declared in-control when it is out-of-control. They use a Markov chain to jointly determine the production cycle, process inspection intervals, and maintenance level. Wang (2006) derives some structural properties for the optimal production/preventive maintenance policy, under the assumption that the (sufficient) conditions for the optimality of the equal-interval PM schedule hold. This increases the efficiency of the solution procedure.

The quality characteristics of the product in a production process can be monitored by a $\overline{x}$-control chart. The economic design of the $\overline{x}$-control chart determines the sample size n, sampling interval h, and the control limit coefficient k such that the total cost is minimized.

Rahim (1994) develops an economic model for joint determination of production quantity, inspection schedule and control chart design for a production process which is subject to a non-Markovian random shock. In their model it is assumed that the in-control period follows a general probability distribution with an increasing failure rate and that production ceases only if the process is found to be out of control during inspection. However, if the alarm turns out to be false the time for searching an assignable cause is assumed to be zero. Rahim and Ben-Daya (1998) generalize the model of Rahim (1994) by assuming that the production stops for a fixed amount of time not only for a true alarm, but also whenever there is a false alarm during the in-control state. Rahim and Ben-Daya (2001) further extend the model of Rahim (1994) by looking at the effect of deteriorating products and a deteriorating production process on the optimal production quantity, inspection schedule and control chart design parameters. The deterioration times for both product and process are assumed to follow Weibull distributions. It is assumed that the process is stopped either at failure or at the m-th inspection interval, whichever occurs first. Furthermore, the inventory is depleted to zero before a new cycle starts.

Tagaras (1988) develops an economic model that incorporates both process control and maintenance policies, and simultaneously optimizes their design parameters. Lam and Rahim (2002) present an integrated model for joint determination of economic design of $\overline{x}$-control charts, economic production quantity, production run length and maintenance schedules for a deteriorating production system. In the model of Ben-Daya and Makhdoum (1998) preventive maintenance activities are also coordinated with quality control inspections, but they are carried out only when a preset threshold of the shift rate of the production process is reached.

2.6.2.2 EMQ problems with failure aspects

A couple of articles study the EMQ model in the presence of random machine breakdowns or random failures of a bottleneck component. For instance, Groenevelt *et al.* (1992a) consider the effects of stochastic machine breakdowns and corrective maintenance on economic lot sizing decisions. Maintenance of the machine is carried out after a failure or after a predetermined time interval, whichever occurs first. They consider two production control policies. Under the first policy when the machine breaks down the interrupted lot is not resumed and a new lot starts only when all available inventory is depleted. In the second policy, production is immediately resumed after a breakdown if the current on hand inventory is below a certain threshold level. They showed that under these policies the optimal lot size increases with the failure rate. Moreover, assuming a constant failure rate and instantaneous repair times the optimal lot sizes are always larger than the EMQ. Nevertheless, Groenevelt *et al.* (1992a) propose to use the EMQ as an approximation to the optimal production lot size. Chung (2003) provides a better approximation to the optimal production lot size. Groenevelt *et al.* (1992b) study the problem of selecting the economic lot size for an unreliable manufacturing facility with a constant failure rate and general distributed repair times. The quantity of the safety stock that is used when the machine is being repaired is derived based on the managerially prescribed service level.

Makis and Fung (1995) present a model for joint determination of the lot size, inspection interval and preventive replacement time for a production facility that is subject to random failure. The time that the process stays in the in-control state is exponentially distributed and once the process is in out-of-control state, a certain percentage of the items produced is defective or qualitatively not acceptable. Periodic inspections are done to review the production process. The time to machine failure is assumed to be a generally distributed random variable. Preventive replacement of the production facility is based on operation time, *i.e.* after a certain number of production runs the production facility is replaced.

Some other articles are concerned with PM policies for EMQ models. For instance, in Srinivasan and Lee (1996) an (S, s) policy is considered, *i.e.* as soon as the inventory level reaches S, a preventive maintenance operation is initiated and the machine becomes as good as new. After the preventive maintenance operation, production resumes as soon as the inventory level drops down to or below a prespecified value, s, and the facility continues to produce items until the inventory level is raised back to S. If the facility breaks down during operation, it is minimally repaired and put back into commission.

Okamura *et al.* (2001) generalize the model of Srinivasan and Lee (1996) by assuming that the demand as well as the production process is a continuous-time renewal counting process. Furthermore, they suppose that machine breakdown occurs according to a non-homogeneous Poisson process. In Lee and Srinivasan (2001) the demand and production rates are considered constant and a production run begins as soon as the inventory drops to zero. If the facility fails during operation, it is assumed to be repaired, but restoring the facility only to the condition it was in before the failure. Lee and Srinivasan (2001) consider an (S, N) policy, where the control variable N specifies the number of production cycles the machine should go through before it is set aside for preventive maintenance overhaul, which restores the facility to its original condition.

Recently, Lin and Gong (2006) determined the effect of breakdowns on the decision of optimal production uptime for items subject to exponential deterioration under a no-resumption policy. Under this policy, a production run is executed for a predetermined period of time, provided that no machine breakdown has occurred in this period. Otherwise, the production run is immediately aborted. The inventories are built up gradually during the production uptime and a new production run starts only when all on-hand inventories are depleted. If a breakdown occurs then corrective maintenance is carried out and this takes a fixed amount of time. If the inventory build-up during the production uptime is not enough to meet the demand during the entire period of the corrective maintenance, shortages (lost sales) will occur. Maintenance restores the production system to the same initial working conditions.

2.6.3 Deteriorating production system with buffer capacity

In order to reduce the negative effect of a machine breakdown on the production process, a buffer inventory may be built up during the production uptime (as it is done in the EMQ model). The role of this buffer inventory is that if an unexpected failure of the installation occurs then this inventory is used to satisfy the demand during the period that corrective maintenance is carried out.

One of the earliest works on this subject is Van der Duyn Schouten and Vanneste (1995). In their model the demand rate is constant and equal to d (units/time). As long as the fixed buffer capacity (K) is not reached the installation operates at a constant rate of p units/time ($p > d$) and the excess output is stored in the buffer. When the buffer is full, the installation reduces its speed from p to d. Upon failure corrective maintenance starts and the installation becomes as good as new. It is possible to perform preventive maintenance, which takes less time than repair and it also brings the installation back into the as-good-as-new condition. The decision to start a preventive maintenance action is not only based on the condition of the installation, but also on the level of the buffer. The criterion is to minimize the average inventory level and the average number of backorders. Since the optimal policy is difficult to implement, the authors develop suboptimal (n, N, k) control-limit policies. Under this policy if the buffer is full, preventive maintenance is undertaken at age n. If the buffer is not full, but it has at least k items, preventive maintenance is undertaken at age N. Maintenance is never performed unless the system has at least k items. The objective is to obtain the best values for n, N and k.

Iravani and Duenyas (2002) extend the above model by assuming a stochastic demand and production process. Demand that cannot be met from the inventory is lost and a penalty is incurred. Moreover, it is assumed that the production characteristics of the system change with usage and the more the system deteriorates the more its production rate decreases and the more its maintenance operation becomes time-consuming and costly. In a recent article, Yao *et al.* (2005) assume that the production system can produce at any rate from 0 (idle) to its maximal rate if it is in working state. Upon failure corrective maintenance is performed immediately to restore the system to the working state. Preventive maintenance actions can be performed as well. Both the failure process and the times to complete corrective/preventive maintenance is assumed to be stochastic. Thus, in addition to the direct cost of performing corrective/preventive maintenance the non-negligible maintenance completion time leads to an indirect cost of lost production capacity due to system unavailability.

Kyriakidis and Dimitrakos (2006) study an infinite-state generalization of Van der Duyn Schouten and Vanneste (1995). The deterioration process of the installation is considered nonstationary, *i.e.* the transition probabilities depend not only on the working conditions of the installation but on its age and buffer level as well. Furthermore, the cost structure is more general than in Van der Duyn Schouten and Vanneste (1995), since it includes operating and maintenance costs of the installation as well as storage and shortage costs. It is assumed that the operating costs of the installation depend on both the working condition and the age of the installation.

Another way of maintaining the buffer inventory is according to an (S, s) policy, *i.e.* the system stops production when the buffer inventory reaches S and the production restarts when the inventory drops to s. This idea is used by Das and Sarkar (1999). They assume that exogenous demand for the product arrives according to a Poisson process. Back-orders are not allowed. The unit production time, the time between failures, and the repair and maintenance times are assumed to have general probability distributions. Preventive maintenance decisions are made only at the time that the buffer inventory reaches S, and they depend on both the current inventory level and the number of items produced since the last repair/maintenance operation. The objective is to determine when to perform preventive maintenance on the system in order to improve the system performance.

A different approach of dealing with integrated maintenance/production scheduling with buffer capacity is presented in Chelbi and Ait-Kadi (2004). They assume the preventive maintenance actions are regularly (after each T time periods) performed and the duration of corrective and preventive maintenance actions is random. The proposed strategy consists in building up a buffer stock whose size S covers at least the average consumption during the repair periods following breakdowns within the period of length T. When the production unit has to be stopped to undertake the planned preventive maintenance actions, a certain level of buffer stock must still be available in order to avoid stoppage of the subsequent assembly line. The two decision variables are: the period T at which preventive maintenance must be performed, and the level S of the buffer stock.

A recent article of Kenne *et al.* (2007) considers the effects of both preventive maintenance policies and machine age on optimal safety stock levels. Significant stock levels, as the machine age increases, hedge against more frequent random failures. The objective of the study is to determine when to perform preventive maintenance on the machine and to find the level of the safety stock to be maintained.

2.6.4 Production and maintenance rate optimization

An integrated production and maintenance planning can also be made by optimizing the production and maintenance rates of the machines under consideration. In this line of research we mention the work of Gharbi and Kenne (2000), Kenne and Boukas (2003), Kenne et al. (2003) and Gharbi and Kenne (2005). In these articles a multiple-identical-machine manufacturing system with random breakdowns, repairs and preventive maintenance activities is studied. The objective of the control problem is to find the production and the preventive maintenance rates of the machines so as to minimize the total cost of inventory/backlog, repair and preventive maintenance.

2.6.5 Miscellaneous

Finally, we list some articles that deal with integrated maintenance and production planning, but their approaches for modelling or the problem settings are different from the articles in the previous categories discussed earlier. For instance, the model presented in Ashayeri *et al.* (1996) deals with scheduling of production and preventive maintenance jobs on multiple production lines, where each line has one bottleneck machine. The model indicates whether or not to produce a certain item in a certain period on a certain production line.

In Kianfar (2005) the manufacturing system is composed of one machine that produces a single product. The failure rate of the machine is a function of its age. The demand of the manufacturing product is time-dependent and its rate depends on the level of advertisement of the product. The objective is to maximize the expected discounted total profit of the firm over an infinite time horizon.

Sarper (1993) considers the following problem. Given a fixed repair/maintenance capacity, how many of each of the low demand large items (LDLIs) should be started so that there are no incomplete jobs at the end of the production period? The goal is to ensure that the portion of the total demand started will be completed regardless of the amount by which some machines may stay idle due to insufficient work. A mixed-integer model is presented to determine what portion of the demand for each LDLI type should be rejected as lost sales so that the remaining portion can be finished completely.

2.7 Trends & Open areas

Initial publications on models in the production and maintenance area date from the end of the eighties (Lee and Rosenblatt (1987)). Since that time many papers have been published with the majority dating from the 1990s and the new millennium. This may be due to a search bias, because those papers are most likely to be found electronically. The most popular area in this review is also the oldest one, *i.e.* on integrated models for maintenance and production. However, still many papers appear in that area and the models become more and more complex, with more decision parameters and more aspects.

The topics on opportunity maintenance and scheduling maintenance in line with production have also been popular, but maybe more in the past than today. We did expect to find more studies on specific business sectors, but could only find many for the airline sector. That sector seems to be the most popular as it has a lot of interaction between maintenance and production as well as high costs involved. In the other sectors, we do see the interaction, but perhaps more papers will be published in the future. The other sections are interesting, but small in terms of papers published.

In general, the demands on maintenance become higher as public and companies are less likely to accept failures, bad quality products or non-performance. Yet at the same time society's inventory of capital goods is increasing as well as ageing in the western societies. This is very much the case for roads, railways, electric power generation and transport and aircraft. As there are continuous pressures on maintenance budgets we do foresee the need for research supporting maintenance and production decisions, also because decision support software is gaining in popularity and more data become electronically available. A theory is therefore needed in these decision support systems. As several case studies have taught us that practical problems have many complex aspects, there is a high need for more theory that can help us to understand and improve complex maintenance decision making.

2.8 Conclusions

In this chapter we have given an overview of planning models for production and maintenance. These models are classified on the basis of the interaction between maintenance and production.

Firstly, although maintenance is intended to allow production, production is often stopped during maintenance. The question arises when to do maintenance such that production is least affected. In order to answer this question planning models should take into account the needs of production. These needs are business sector specific and thus applications of planning models in different areas have been considered. In comparison with other specific sectors, much work has been done on modelling maintenance in the airline sector. In Chapters 3, 4 and 5 of this thesis we focus on maintenance planning at the railways. In chapter 3 we give an overview of techniques used for planning railway infrastructure maintenance and in chapters 4 and 5 we develop methods for scheduling preventive maintenance activities. In order to reduce costs, downtime and inconvenience for travellers and operators, the preventive maintenance activities are scheduled together as much as possible. Moreover, in chapter 6 we deal with a problem that arises within the passenger railway operator when the timetable is modified due to some extra trains or scheduled maintenance work on some parts of the railway infrastructure. This problem is called the Rolling Stock Rebalancing Problem.

Secondly, maintenance itself can also be seen as a production process which needs to be planned. Models for maintenance production planning mainly address allocation and manpower determination problems.

Finally, maintenance also affects the production process since it takes capacity away. In production processes maintenance is mostly initiated by machine failures or low quality items. Maintenance and production should therefore be planned in an integrated way to deal with these aspects. Indeed, integrated maintenance and production planning models determine optimal lot sizes while taking into account failure and quality aspects. We observe a non-stop attention for such models, which take more and more "real world" aspects into account.

Although many articles have been written on the interaction between production and maintenance, a careful reader will detect several open issues in this review. The theory developed thus far, is far from complete and any real application, is likely to reveal many more open issues.

Chapter 3

An Overview of Railway Maintenance and Track Possession Planning Models

Abstract

In this chapter we give a literature review on the planning of maintenance of railway infrastructure and especially on the effect of maintenance on railway capacity. The review is done on the basis of a subdivision of the planning process in eight phases. It concentrates on two main topics: firstly diagnosing the state of the track and the subsequent planning of the maintenance work and secondly the planning of track possession for maintenance. Several approaches for the latter are outlined.

This chapter is based on Budai and Dekker (2003).

3.1 Introduction

Railways are important transportation means in the European Union policy for sustainable transport. Therefore, in the last decades many changes have been introduced to improve service and efficiency. In order to allow for competition between rail operators on the same track, separate rail infrastructure management (RIM) companies have been introduced, *e.g.* Railinfrabeheer (RIB) in 1995 in the Netherlands and Railtrack in 1994 in the United Kingdom. Railtrack was privatized but the infrastructure management was later in 2001 brought back under government control and put under Network Rail, while RIB was merged in 2003 with the rail traffic control and the rail planning company Railned into the government owned ProRail B.V., which obtained an infrastructure management contract for 10 years. In some countries the actual maintenance has been outsourced to specific contract maintenance operators. Although these measures were intended to create better management, actual practice revealed several shortcomings. Hence it has become clear that maintenance management is substantially more difficult than thought, and that scientific literature on this subject shows substantial gaps.

These research needs have been acknowledged by the EU in their research programs. One of the projects initiated is IMPROVERAIL (see Improverail (2002)), which deals with defining the overall state of the art in railway infrastructure management. Part of this project is aimed at making an overview of railway infrastructure maintenance planning and then to determine best practices. This chapter gives an overview of the scientific literature on infrastructure maintenance, excluding the rolling stock from our research field.

Although good decisions require good information, we exclude information systems from our discussion. Our practical experience indicates that information in railway infrastructure companies tends to be scattered over many databases and spreadsheets and that a central common database with information over the entire railway infrastructure and its condition would be a significant step forward. Yet the development of such an install-base database is an enormous project for most railways, as their infrastructure has been built up over a century and proper documentation has not always been maintained.

The review we present in this chapter is based on a general structuring of maintenance planning. It also applies to other types of maintenance, like road, distribution systems (electricity, water, *etc.*) and other civil maintenance (see *e.g.* Dekker (1996) for an overview of maintenance optimization models and Dekker *et al.* (1998a) for highway maintenance). The ideas behind this structuring came forward while developing a decision support system (DSS) for opportunity maintenance at gas turbines (see the PROMPT project in Dekker (1996). Little has been written about structuring maintenance planning as in fact most maintenance research so far has focused on the link between maintenance and production (see *e.g.* Budai *et al.* (2008)) and on determining the maintenance concept (see *e.g.* Pintelon *et al.* (1997)). Other examples for the latter category are the EUT model from the Eindhoven group (Geraerds (1991) and Gits (1984)) and Kelly's model (see Kelly (1989)). Next there is much literature on improving a concept once experience is gained (delay time models, see Christer and Waller (1984) and condition monitoring models).

We start this chapter with a presentation of the structuring of the maintenance planning process. Next we present the literature overview. We conclude this chapter with some future research topics and final conclusions.

3.2 A structuring of maintenance decisions at railways

We first give some general definitions. Maintenance can be defined as the process to keep a system in a state where it can perform its function as desired. Maintenance consists of inspections, first-line maintenance, repairs, replacement of parts or modules and renewals of whole systems. Maintenance can be triggered by failures or be done before failures, *i.e.* preventively. In the latter case it can be triggered by the passing of time or by condition indicators, in which case it is called condition-based maintenance. In the planning a further distinction is necessary: maintenance can either be routine or a project. A project is a substantial task of work, which is planned beforehand and carried out once every few years, while routine maintenance is the small work done repeatedly. Each type has different planning characteristics.

3.2.1 A general structuring of the maintenance planning process

The structuring of the maintenance planning process, which is presented below, is a top-level breakdown of operations to define the maintenance work to be carried out where and when. Apart from presenting the structuring we also briefly explain how planning is done with some infrastructure managers. We often take the case of the Netherlands, with ProRail as infrastructure manager, as a benchmark. In the IMPROVERAIL project the Dutch situation is compared with that of other European infrastructure managers.

In structuring of the maintenance planning process we distinguish the following decision phases.

1. Budget determination
2. Long-term quality prediction
3. Project identification & definition (diagnosis)
4. Project prioritization and selection
5. Possession allocation and timetabling of track possession
6. Project combination
7. Short term maintenance and project scheduling
8. Work evaluation and feedback loop

The structuring is based on the authors' own experience with maintenance planning at some companies. The year cycle is a main driver behind the structuring as it corresponds to the yearly budget cycle and the project prioritisation. In practice phases can be much more complex than described here and they are not always done sequentially, but sometimes repeated in cycles or smaller steps.

3.2.2 The various decision processes

Phase 1 - Budget determination

Most RIMs receive their budget from the government as railway infrastructure is considered to be a public good. Governments and many other organizations fix their budget annually. Often the budget may be translated into budgets for regions or individual projects. Usually an overall budget is set for routine maintenance and for the many small projects. Separate budgets may be defined for major new construction projects, which improve the functionality of the railway system.

Two basic methods exist to determine budgets. The first is based on history; last years budget is taken as basis and adapted for inflation, efficiency gains, special events, *etc.* Although this widely used method lacks a link with the actual operations and state of the infrastructure, the underlying idea is that the most crucial shortcomings will come forward in time and get extra funding. However, the reality is that both in the UK and the Netherlands, the RIMs recently got major increases of budget, because of the recognition and acceptance of severe deterioration in the infrastructure caused by budget cuts in previous years. The second method consists of a breakdown of the whole infrastructure into elements. Next for each element either a desired quality level is set, from which a maintenance policy is derived, or a (optimal) maintenance policy is determined from a cost-optimisation. The resulting maintenance cost is then summed over all elements and this gives the annual budget. The problem with this approach is that all infrastructure elements should be taken into account and that quality norms should be set, which is a substantial task.

Phase 2 - Long-term quality prediction

Closely coupled to the previous phase is the prediction of the quality of the infrastructure on the long-term. The inputs for this process are the continuous inspections carried out on the infrastructure. Besides that there is the prediction of the railway capacity needed in the future. Once the budget is determined, one can determine how much maintenance can be carried out and what will be the evolution of the quality of infrastructure in the long-term. This requires predictive deterioration models, which indicate the deterioration over time conditional on maintenance interventions. If the long-term quality level is below an acceptable norm, budget increases may be requested.

Phase 3 - Project identification & definition (diagnosis)

Maintenance is split into routine maintenance and projects. Inputs in this process phase are the maintenance concept and the outcomes of phase 2, which clarify where maintenance and renewal (M&R) work is needed. Next one decides how it will be renewed and which other parts (not yet needing maintenance) will be done as well. This project definition seems to be a manual process within many RIMs. The required track possession will also be determined here (in terms of hours needed, presumably with day or night work options, but dates remain open). Within several RIMs projects are defined on a five-year horizon (*e.g.* ProRail, Swiss Railways, *etc*).

Phase 4 - Project prioritization and selection

In the previous phase, projects are defined for the coming years. Next one has to determine which projects will be carried out when. This can be done by using priority indices for each project (what is their contribution to the RIM goals) and then selecting the project with the highest priority. In the Netherlands the following objectives are used: reliability, availability, safety, comfort and environment. Each project is scored on its contribution to each objective. Next a multi-criteria decision method is used to determine an overall priority criterion. The final selection of the projects is then based on the available budget and on the balance of the workload for the contractors carrying out the projects.

Phase 5 - Possession allocation and timetabling of track possession

Once the projects have been selected, preparation starts in detail. We will not go into detail regarding how each project will be planned, as that is a separate topic. Apart from tendering the projects, one also determines in which way the projects will be carried out if they require track possession. Several alternatives exist, each with different consequences for the total costs. Some work may be done during train service, implying only train speed reduction and with a danger of accidents. Recently, this way of operation was banned in the Netherlands after some tragic accidents. One may do project work at night when there is little train traffic (except for some freight trains). Night work however, is also limited, as also time is needed to shut down the track and some buffer time to allow for late trains. This also means that equipment has to be moved in place and removed before the standard train service starts again. An alternative is to close down the track, *e.g.* in a weekend or during a low-demand period, like the summer. In that case alternative transport can be arranged, such as buses. Depending on the situation, this may or may not pose problems. Anyhow, maintenance productivity will be much higher if one can work during a longer time period. The possession planning usually is done long beforehand, *e.g.* more than a year beforehand in the Netherlands. A complicating factor in the UK is that several operators are using the track, with different possession priorities, which may veto possession. This has led to too little time for maintenance in the UK for NETWORK RAIL.

Phase 6 - Project combination

This yearly phase takes as starting point the list of agreed projects for the coming year. Next it determines whether these projects can be combined with other work, *e.g.* routine maintenance or some not yet approved works, in order to reduce track possession time or to save execution costs. Some part of this work is already done in phase 4, when a project is defined.

Phase 7 - Short-term maintenance scheduling

Finally, given the project selected for the coming year and the routine maintenance to be carried out, a time planning is made, specifying when each task will be carried out. In this plan the capacity to do the work in terms of manpower and specific equipment needed is balanced over the year, while the track possession restrictions are taken into account as well.

Phase 8 - Work evaluation and feedback loop

After M&R work completion, the effects on the quality of the infrastructure are assessed. Several performance criteria can be used in this respect, such as time period measures of the number of failures, incidents, the amount of corrective maintenance, *etc.*

3.2.3 Typical aspects for railway maintenance

The structuring presented is quite general and similar methods can be used in different business areas. Specific to railways is the long planning cycle and the way the possession is determined in phase 5, as this requires input from the train timetable. In general one has to fit the downtime required for maintenance execution within the time needed for production. If production is halted in the weekends one can then do maintenance. Railway operation is however, like a continuous process industry. In the petrochemical industry one often uses yearly or four yearly shutdowns to do maintenance. In the case of railways the balancing of maintenance and production is very difficult, because the production is already specified in the yearly timetable and skipping a connection completely overthrows the equipment rotation. It does have some resemblances with the planning of maintenance of highways, although cars do have more flexibility than trains (see *e.g.* Dekker *et al.* (1998a)).

In the next section we will present an overview of published papers concerning these planning phases with railways.

3.3 Literature overview

The literature review was carried out by searching through (online) databases, like ScienceDirect, JStor, MathSciNet, Google Scholar, Scirus and Scopus, by using personal contacts with railway researchers and by following references from papers retrieved. Some fifteen papers were found. We will discuss them following the structuring provided before.

No contributions were found on phase 1, the budgeting process, on phase 3, the project prioritisation, on phase 6, project combination and phase 8, work evaluation. These are general activities, which do not have railway specific elements, though they are a major issue in the railways. The papers found can be split into two groups. The first group incorporates a deterioration modelling (phase 2), defines the resulting work (phase 3) and makes some scheduling. The second group takes the maintenance work as input and tries to fit it in the timetable and makes a good manpower and equipment schedule. Below we discuss contributions from each group.

3.3.1 Group one - papers with deterioration modelling

In total thirteen papers were identified. The first eight are stand-alone papers by Miwa *et al.* (2001), Levi (2001), Simson *et al.* (1999), Ferreira and Murray (1997a), Meier-Hirmer *et al.* (2005), Hokstad *et al.* (2005), Podofillini *et al.* (2006) and Lamson *et al.* (1983). Next there are two papers by Jovanovic (see Jovanovic and Korpanec (2000) and Jovanovic and Zaalberg (2000)) on ECOTRACK, a decision support system for railway track maintenance. Finally, there are three papers by Zoeteman (see Zoeteman (2001), Zoeteman and van Zelm (2001) and Zoeteman and Braaksma (2001)), who focus on life cycle costing and discuss the scheduling of maintenance only briefly. Below we discuss the papers one by one.

In the paper of Miwa *et al.* (2001) the authors developed a mathematical programming model for the optimal decision-making for tie (or sleepers)[1] tamping. The model is built in two steps. Firstly, the authors have built a transition model for predicting changes in the surface irregularity and then a mathematical programming model is applied for making optimal maintenance strategies for the annual tamping schedule. The tamping is done

[1]A detailed description of the terminology is given in Appendix A.

with a Multiple Tie Tamper (MTT) machine that is shared by several track maintenance depots. The model indicates which month the MTT should be allocated to a particular depot and which lot (1 lot = 100m) should be provided with maintenance work.

Renewal differs from maintenance work in terms of scale and frequency. Components, which are maintained to a higher standard, require a greater level of maintenance input, but are likely to require less frequent renewal. Conversely, a reduced level of maintenance will generally cause assets to degrade more quickly, leading to a more frequent requirement for renewal. What is the optimal balance between renewal and maintenance? The paper of Levi (2001) gives an answer to this question by developing specific software for optimization, based on a maintenance model. The optimization model can be used for the evaluation of maintenance costs and their dependence on traffic, and for spot renewal decisions. This paper presents a number of issues to be addressed within this task, such as: modeling reliability, modeling maintenance and renewal costs, and optimizing the decision in acceptable time. He assumes that the maintenance costs can be split into preventive maintenance (depending on traffic and increasing in time) and repair. Further assumption is that the renewal decision does not depend on maintenance costs, but on the increase of repair costs. If there were no renewal at all, the components would be in time entirely replaced piece by piece. The way to calculate the amount of repairs by time and track length unit is presented in the article.

Nowadays the trend is to increase the pressure on rail track infrastructure in terms of increases in axle loads and train speeds. These lead to reductions in the life of track components and increases in maintenance works and costs. An important issue therefore is the modeling of the track degradation. This issue is briefly presented in Simson *et al.* (1999), Ferreira and Murray (1997a), Meier-Hirmer *et al.* (2005), Podofillini *et al.* (2006) and Hokstad *et al.* (2005).

Simson *et al.* (1999) presents a model that provides a simulation of the costs and benefits of degrading railway track conditions and by the maintenance work that may be conducted. Ferreira and Murray (1997a), besides track degradation modelling, also gives an overview of track maintenance decision support systems currently used in North America and Europe. The paper focuses on three important aspects, which need to be considered when decision support tools for efficient maintenance planning are developed. These are: the physical factors which affect deterioration and therefore costs of rectification or renewal, the scope and capabilities of existing track degradation and maintenance planning models, and the parameters, which must be included in the optimization processes to take into account engineering and business related factors.

Meier-Hirmer *et al.* (2005) present the way in which the track maintenance optimization is carried out at the French national railway company, SNCF. The track geometry measurements are recorded and based on the statistical analysis of the available measurements probabilistic models of the degradation and intervention efficiency are derived.

These models are then combined in a global model of track maintenance costs, which are optimized according to the inspection steps and intervention levels. Hokstad *et al.* (2005) focuses on modelling the failure mechanism of a specific railway line in Norway. They consider two main failure mechanisms, either an immediate critical failure or a critical failure that occurs as the result of a degradation process. Various types of inspection and maintenance actions are performed on the line. The degradation/repair process is modelled as a time continuous Markov chain. Podofillini *et al.* (2006) develops a methodology for the determination of an optimal inspection/maintenance strategy with the objective of reducing the operation and maintenance costs, but still assuring high safety standards. First a risk/cost model has been built accounting for the realistic issues of the rail failure process and for the actual inspection and maintenance procedures followed by the Norwegian railway company. Then a multi-objective optimisation approach has been developed to optimise inspection and maintenance procedures with respect to both economical and safety related aspects. The latter optimisation step is done by using genetic algorithms.

Two studies concerning the maintenance and replacement of heavy haul railway track are presented in Lamson *et al.* (1983). The authors consider in both studies rail in a heavy haul railway line, which is used to transport metal ore from mine to port. It is assumed that the wear mode in curved track is different from that in straight track; the curved track necessitates much more frequent replacement than straight track. The first study demonstrates the application of the decision network method in determining minimum cost maintenance and replacement policies for curved rails in heavy haul rail track. Moreover, a detailed sample calculation for a typical curve class (rail wear and defect characteristics are similar for rails within the same class) is presented. The second study develops a computer model, which optimises the time for straight rail replacement, using their fatigue defect distribution functions. The incorporation of taxation and financial factors is an important requirement of this optimisation technique, since the tax rates on rail replacement cost (capital expenditure) and rail maintenance cost (operating expenditure) are different. In addition, the financial factors (return rate, inflation rate, escalation rate on materials and labour) are directly related to the economics of the rail management decision.

Achieving efficient infrastructure can be accomplished with proper maintenance management helping the decision-making process that would be directed towards maximum track safety and reliability. If we take into account the size of the railway networks and the complexity of the relationship between the various parameters and their joint or separate influence on the track infrastructure component condition, we can see that there is a huge amount of data that has to be handled and processed in order to reach a diagnosis. These facts show the necessity of having computer-aided decision support systems (DSS) and these were the reasons behind the development of ECOTRACK software. ECOTRACK (see Jovanovic and Korpanec (2000) and Jovanovic and Zaalberg (2000)) is one of the major software programs for track maintenance and renewal in Europe. It provides a solution to the problem of maintaining track at the required quality level for minimum cost, which is especially problematic in the new environment of increasing loads and speeds and changes in responsibilities. For other decision support and computerized maintenance management systems developed for and used in the European rail sector we refer to Zoeteman (2006) and Carretero *et al.* (2003).

ECOTRACK is structured as a five level system, each level representing a planning activity. The first three steps are based on an analysis of the track condition with an increasing level of detail. In the initial diagnosis the rough M&R needs are determined per track component. The work plan is refined in the detailed diagnostic, where additional information is collected from the engineers. The preliminary work plan is improved by clustering works of the same (or different) type if they fulfill some requirements (time and location). The fourth level enables the user to carry out optimization when the available resources are limited. Finally, the fifth level provides a set of tools required by the planner to achieve optimal management of M&R work on the whole network. ECOTRACK's rules are based on all the key data, as track geometry, the condition of track components and the effects of the different M&R activities and methods. This software does not do a possession planning and only considers track and no other infrastructure elements, such as switches or catenary system.

Zoeteman and Braaksma (2001) deal with the issue of improving the long-term performance of railway system, more precisely with developing a DSS for evaluating track design of the Dutch High Speed Line. The High Speed Line South (HSL South) is the Dutch part of the high-speed link from Amsterdam to Brussels, Paris and London. New tracks, with a total length of almost a hundred kilometers, will be designed for 300 kph. With the DSS construction, maintenance, renewal, failure and financing data are combined to make estimates of total life cycle costs. The infrastructure performance (availability and reliability) is included in these estimates, since it influences the costs and revenues of transport operations.

Although there are only few papers, they do show that advanced techniques are needed and that they are successful in optimising railway infrastructure maintenance. The ECOTRACK system is a very advanced system, which took years to develop. Although it is a major step forward, it also shows that more research is needed to come to an overall optimisation system.

3.3.2 Group two - papers concentrating on maintenance scheduling in line with track possession

Six papers were found which focus on track possession planning and one paper (Grimes (1995)) that deals only with the maintenance scheduling problem. In Higgins (1998) and

Ferreira (1997b) the track possession is modelled in between operation. This can be done for occasionally used track, which is the case in Australia. If tracks are used frequently, one has to go over to night/weekend maintenance, as in Budai *et al.* (2006). In that case one can either make a cyclic static schedule, which is made by Den Hertog *et al.* (2005) and Van Zante-de Fokkert *et al.* (2007) for the Dutch situation or a dynamic schedule with a rolling horizon, which has to be made regularly. The latter is described by Cheung *et al.* (1999). Below we discuss all contributions in detail.

Higgins (1998) presents an optimization model for determining the best allocation of railway maintenance activities and crews in order to minimize the disruption to and from scheduled trains. It minimizes the amount of time a given track segment has a service level below a specified benchmark (this is achieved by scheduling each of the activities as early as possible). Both trains and maintenance activities are subject to unexpected delays. A late train can be delayed further or it can delay the start of a maintenance activity, while a delayed activity can disrupt one or more train services. The tabu search heuristic technique was used to find a solution to the model.

Budai *et al.* (2006) introduces the preventive maintenance scheduling problem (PMSP), where (short) routine activities and (long) unique projects have to be scheduled in a certain period. In order to reduce costs and inconveniences for the train operators and travelers, these maintenance activities should be scheduled together as much as possible. Apart from giving a math programming formulation for the PMSP, the authors present also a couple of fast heuristics.

Another article that deals with the scheduling of maintenance work is written by Van Zante-de Fokkert *et al.* (2007). This paper presents the construction of a four-week cyclic preventive maintenance schedule for the Dutch Railways (only the routine maintenance). All maintenance activities are carried out within working zones and during maintenance these zones are taken out from service exactly once, at night for five and a half hour (for the main lines). The maintenance work of the peripheral railroads and yards are planned in the daytime. In this article only the construction of the nightly maintenance schedule for the main lines is presented. The advantage of having such a fixed schedule for routine maintenance is that it simplifies the overall possession planning considerably since no extra possession for routine maintenance needs to be scheduled and an overall workload balancing can be achieved across the country. The introduction of the concept did require that all maintenance work should be shifted to night hours. This was done on safety grounds.

The first part of a companion article by Den Hertog *et al.* (2005) contains a method of dividing the Dutch Railway system into working zones. A working zone is a part of the railways that is taken out of service, when track workers are working in it. The working zones are indicated on maps and small signs (unique codes) are attached to the rails. The workzones use switches as basis and are based on the fact that if one puts a switch out of operation the whole line segment behind it can no longer be used until the next switch. The method of constructing a maintenance schedule is a two-step method; in the first step the Single Track Grids (STG) are specified and in the second step they are assigned to the nights. STG is a set of working zones that can be taken out of service simultaneously without disrupting severely the nightly railway traffic. The left and the right tracks of an STG are assigned to different nights, because in this way there are no completely blocked corridors. The objective of the model is to minimize the number of nights with planned maintenance in the schedule and the sum of the maximum scheduled workload.

The Railway Track Possession Assignment Problem, which is a resource-allocation problem, appears in the paper of Cheung *et al.* (1999). The problem is to assign railway tracks to a given set of scheduled maintenance tasks according to a set of constraints. One of the requirements of the Railway Track Possession Assignment Problem is to schedule sufficient preventive maintenance work so that the railway can operate without any disruption during the operated period (ca 19h every day). Another requirement is that the time available to carry out maintenance work is less than 5 hours a day. Furthermore, the number of types of maintenance work required is quite high and a set of rules and procedures have to be followed in scheduling the maintenance work (*e.g.* safety rules). Moreover, the maintenance tasks have priorities, so the purpose is to assign as many high-priority job requests as possible.

Another approach, developed by Ferreira (1997b), focuses mainly on the Australian railway system and contains a survey of some planning models for freight rail operations. The survey includes optimization and reliability of train schedules, train dispatching, locomotive/wagon scheduling, rail yard activities, intermodal terminal planning, track capacity and maintenance planning. Among other things, the paper deals with track infrastructure planning, including track capacity and track maintenance planning issues from an Australian perspective. According to the paper the planning for a rail line involves determining the number, length and location of sidings[1] required. In order to determine the position of the sidings, two main assumptions are made, namely: the sidings are located in order to minimize train delays and/or total train operating costs; and the schedule is repeated after a period of time (*e.g.* daily, weekly). Insufficient sidings, or sidings placed at inappropriate locations, may result in poor train performance. Ferreira (1997b) presents also an overview of works for determining the best positions of sidings using simulation.

The focus of Grimes (1995) is not that much on possession planning, but more on planning the track maintenance works. In this paper track tamping is planned using Genetic Algorithm (GA) and Genetic Programming (GP) methods, with profit as the optimization criteria. The results were compared with an existing deterministic technique and it was found that the GP method gave the best results, with the GA method giving good results for short sections of track (10 miles) and poor results for long sections (50 miles).

The five papers show that track possession planning is an important issue, which is driven by safety of the workers and minimal possession to rail operators. The best approach depends on how much the infrastructure is used. As all papers are quite recent we conclude that this is an interesting subject on which more papers can be found once other countries adopt similar approaches. So far, all papers consider only one rail operator. The problem becomes even more complex if multiple operators with different time preferences (*e.g.* passenger versus cargo operators) have to be taken into account. This situation applies to the UK and to some extent to the Netherlands as well.

3.4 Conclusions and further research

The papers found so far only provide a partial solution to railway infrastructure maintenance management. Considering the overall structuring we first notice that for several parts no papers exist. This may be due to a publication bias, because often papers are only written if they contain sufficient academic models and several parts are more qualitative than quantitative. With respect to the phases on which literature has been published we would like to remark that an overall method is lacking. Extensions should be made in the following ways:

- Determine budgets from a model of deterioration of individual components, extending ECOTRACK to other infrastructure elements
- Combine knowledge on infrastructure deterioration with life cycle cost models and with possession planning to produce a good maintenance plan
- Develop methods to combine various projects requiring possession

In Chapters 4 and 5 of this thesis we deal with the last two issues mentioned above. We develop several methods for scheduling preventive maintenance activities consisting of routine works and projects such that the sum of the track possession costs and the maintenance costs is minimized. Each of these works has different planning characteristics, such as frequency, duration, cost, etc. Moreover, the routine works and projects are combined as much as possible in order to minimize the track possession times. In the next two chapters we do not explicitly model the infrastructure deterioration. However, we assume that the maintenance work resulting from deterioration modelling and/or track inspections can be specified as jobs that should be executed within a certain time window or with a certain frequency. Accordingly it can be treated as a project or a routine work in our planning problem.

Chapter 4

Scheduling Preventive Railway Maintenance Activities

Abstract

A railway system needs a substantial amount of maintenance. To prevent unexpected breakdowns as much as possible, preventive maintenance is required. In this chapter we discuss the Preventive Maintenance Scheduling Problem (PMSP), where (short) routine activities and (long) unique projects have to be scheduled in a certain period. To reduce costs and inconvenience for the travellers and operators, these activities should be scheduled together as much as possible. We present two versions of the PMSP, one with fixed intervals between two consecutive executions of the same routine work and one with only a maximum interval. Apart from giving a math programming formulation for the PMSP and for its extension we also present some heuristics. In addition, we compare the performance of these heuristics with the optimal solution using some randomly generated instances.

This chapter is based on Budai, Huisman and Dekker (2006).

4.1 Introduction

Reliability, that is, punctuality and safety, are important aspects in railway transport. The quality of the railway infrastructure has a major influence on the reliability of the railway system as a whole. Therefore, it is important that there is enough preventive maintenance of the infrastructure (*e.g.* rail, ballast, sleepers, switches and fasteners). However, maintenance is expensive and budgets for maintenance are always under pressure. So it is important to reduce the maintenance costs without reducing the maintenance itself. This chapter considers the clustering of maintenance activities on the same link in a network in order to reduce the disturbance of the railway traffic.

We introduce the *Preventive Maintenance Scheduling Problem (PMSP)*, where a schedule for the maintenance activities has to be found for one link such that the sum of the possession costs and the maintenance costs is minimized. The possession costs are mainly determined by the possession time, which is the time that a track is required for maintenance and cannot be used for railway traffic. The focus is on medium term planning, determining which preventive maintenance works will be performed in what time periods (month/week/hours).

We assume that the execution of preventive maintenance is not influenced by corrective maintenance. Hence it is of the block maintenance type. Moreover, we assume that the maintenance frequency has been determined in an earlier stage (in the design or maintenance concept phase). Our model focuses on operational planning. Many scientific papers consider the relation between preventive and corrective maintenance, but these trade-offs are hardly done in the operational planning phase in practice. Hence we do not consider such a trade–off in our work.

The contribution of this chapter is twofold. First, we will provide a math programming formulation for the PMSP and for its extension and we prove that PMSP is NP-hard. Second, we will give four heuristics to solve this problem quickly. The latter is important, since in practice a whole network needs to be optimized. Moreover, to give an insight into the quality of the heuristics, we will compare them with (a lower bound on) the optimal solution.

Although we have formulated our problem for railway maintenance, the maintenance scheduling problem also arises in other public/private sectors, since preventive maintenance of other technical systems (machine, road, bridge, building, high voltage lines, electric power stations, airplanes) contains also small routine works and large projects (*e.g.* see Kralj and Petrovic (1988)). Several papers have addressed the grouping of preventive maintenance activities in general. For instance, Van Dijkhuizen and van Harten (1997) consider a clustering problem for frequency-constrained maintenance jobs with common and shared set-ups in an infinite horizon setting. They only consider cyclic policies, while we also allow non cyclic policies. Wildeman *et al.* (1997) present a rolling horizon approach to group maintenance activities on a short-term basis. In Dekker *et al.* (2000) a general approach for the coordination of maintenance frequencies is presented. In both papers there are no repetitive jobs, however, and a continuous-time approach is taken. The maintenance scheduling problem under a deterministic environment for a group of non-identical machines is studied by Hariga (1994). Here only cyclic overhaul schedules are considered and it is assumed among others that for each machine the cycle time of the major overhauls is an integer multiple of the minor overhaul intervals. Another work relevant to maintenance scheduling is that of Sriskandarajah *et al.* (1998) in which an approach for scheduling the frequency-based overhaul maintenance is given. The objective here is to satisfy the maintenance requirements of various train units as closely as possible to their due dates, since there is a cost for undertaking the maintenance tasks too early or too late.

The contribution of our study to the existing literature on maintenance optimization is that we consider both repetitive jobs and projects in a finite horizon with a small (weekly) time step instead of infinite horizon or problems with projects only. This more closely resembles practical maintenance scheduling problems.

The remainder of this chapter is organized as follows. In the next section we place our study in the railway world and the existing literature on maintenance scheduling of railway infrastructure. A formal problem description is given subsequently, followed by a mathematical programming formulation for PMSP and its extension. After that we present some heuristics to solve this problem approximately. Finally, we conclude this chapter with some computational experiments.

4.2 Railway maintenance planning

Preventive maintenance on railways can be subdivided into small routine works and projects. The routine (spot) maintenance activities consist of inspections and small repairs, for example, inspection of rails, switch, level crossing, overhead wire, signalling system and switch lubrication (see Esveld (2001)). These works do not take much time to be performed and are done regularly, with frequencies varying between monthly and annual. The projects include renewal works and consist, for example, of ballast cleaning, rail grinding and tamping (see Esveld (2001)). They are only carried out once or twice every few years.

IMPROVERAIL (see Improverail (2002)) and Budai and Dekker (2002) show that the preventive railway maintenance works are carried out in most countries during train service. In the actual train timetable possible possession allocations are scheduled for maintenance so that it should not affect regular train services too much. Some methods are presented in Higgins (1998) and Cheung *et al.* (1999). Many countries use timetabling software for finding free intervals or periods with less impact to the train operators. Carrying out maintenance works during train services might be unsafe for the maintenance crew. Therefore, in some other countries (*e.g.* The Netherlands) the maintenance works are carried out either during night (when there are only few or no trains) or during the day with interruption of the train service. In the first case one can make a cyclic static schedule, which is made by Den Hertog *et al.* (2005) and Van Zante-de Fokkert *et al.* (2007) for the Dutch situation or a dynamic schedule with a rolling horizon, which is presented in Cheung *et al.* (1999). Miwa *et al.* (2001) present an optimal schedule for only one type of maintenance work, namely an annual schedule for the tie (sleeper) tamping.

Although our approach considers one rail link only, it can be extended to a network using the concept of the Single Track Grids (STGs) presented in Van Zante-de Fokkert *et al.* (2007). Solving the preventive railway maintenance scheduling problems to optimality involves large scale mixed integer programming problems. Several heuristics are developed to solve these problems faster, for example, Higgins (1998) applies tabu search techniques, while Grimes (1995) uses genetic algorithm and genetic programming methods.

4.3 Problem description

In this chapter our aim is to give a schedule for preventive maintenance activities in a finite horizon, such that jobs are clustered as much as possible in the same period. Combining jobs as much as possible results in costs savings, since execution of a group of activities requires only one track possession.

The PMSP can be defined as follows. Given a set of routine activities and projects, we like to schedule them such that the track possession costs and the maintenance costs are minimized. Some routine works and projects may be combined to reduce the possession time, but others may exclude each other. For each routine work the planning cycle, that is, the maximum number of time periods between two consecutive executions, is known. Furthermore, the number of time periods elapsed since the routine works have been carried out for the last time is given. A list of projects that need to be performed in the planning period, duration and the earliest and latest possible starting times for each project are given. The execution costs of each routine work and project and the costs of having a track possession in the planning period are known. A list of works (routine works and/or projects) that can be combined is given. Since our planning horizon is finite and the routine activities are repetitive, an end-of-horizon valuation is needed.

At first sight, the model presented in this chapter seems to be related to the machine scheduling problem and the multi-project scheduling problem in an abstract way. The similarity among these three problems is that there are jobs, with given durations and given time windows between two consecutive executions, which have to be scheduled in a certain time period. One of the differences is that in the PMSP the routine works have a repetitive character. Furthermore, the objective of the PMSP is different from the objectives of the other mentioned problems. Namely, we try to schedule the jobs together as much as possible and not necessarily as soon as possible.

Note that we do not take corrective maintenance work into account in the planning. We assume that track possession for this work is arranged in a separate manner if necessary. As the nature of corrective work is stochastic and it does not occur that often, it is difficult to incorporate it in a planning and track possessions need to be agreed and communicated long beforehand. We also do not explicitly address condition-based maintenance. Several infrastructure companies have special railcars for inspections, such as the Eurailscout Inspection and Analysis, which regularly check the tracks on defects and quality problems. In the Netherlands such a check is carried out four times a year. We assume that the maintenance work resulting from the inspections can be specified as jobs to be executed within a certain time window (not all work is urgent and it may require some preparation). Accordingly it can be treated as a project in our planning problem. We do acknowledge that due to these random events the planning may have to be redone, but that is not uncommon in planning processes. In fact, a frequent replanning necessitates the use of a computerized planning tool.

4.4 Mathematical Formulation

Let T be a set of discrete time periods (*e.g.* months, weeks) in which the maintenance activities need to be scheduled, *i.e.* $|T|$ is the planning horizon.
The model parameters are as follows:

PA a set of projects,

RA set of routine maintenance works,

A $PA \cup RA$ set of all activities,

C $\{(m,n)|$ work m is combinable with n, $\forall m,n \in A\}$ set of combinable works,

L^a cycle length of the routine work $a \in RA$,

F^a $\lfloor \frac{|T|}{L^a} \rfloor$ frequency of the routine work $a \in RA$,

G^a number of periods elapsed since routine work $a \in RA$ was in the past (before the planning horizon starts) for the last time carried out,

LC^a $\{t \in T | 1 + |T| - L^a \leq t \leq |T|\} \subseteq T$ set of time periods from the last planning cycle for routine work $a \in RA$,

b^a_t $\frac{(|T|-t)}{L^a}$ length of the remaining interval until the end of planning horizon divided by the length of the planning cycle for routine work $a \in RA$ and for time period $t \in LC^a$,

$T_p \subseteq T$ set of possible start points of project $p \in PA$,

D_p duration of project $p \in PA$,

pc_t possession cost in period $t \in T$,

mc_a maintenance cost per time period for carrying out work $a \in A$.

The following binary decision variables are defined:

x_t^a binary variable that denotes whether activity $a \in A$ is assigned to period $t \in T$ ($x_t^a = 1$), or not ($x_t^a = 0$),

z_t^a binary variable that denotes whether activity $a \in RA$ is carried out for the last time in the planning horizon at time $t \in LC^a$ ($z_t^a = 1$), or not ($z_t^a = 0$),

m_t binary variable that denotes whether the track is used for preventive maintenance work at time $t \in T$ ($m_t = 1$), or not ($m_t = 0$),

y_t^p binary variable that denotes whether the execution of project $p \in PA$ starts at time $t \in T$ ($y_t^p = 1$), or not ($y_t^p = 0$).

The PMSP can now be formulated as follows:

$$(PMSP) \quad Min \sum_{t \in T} pc_t m_t + \sum_{a \in A} \sum_{t \in T} mc_a x_t^a + \sum_{a \in RA} \sum_{t \in LC^a} mc_a b_t^a z_t^a \tag{4.1}$$

s.t.

$$\sum_{t=1}^{L^a - G^a} x_t^a \geq 1 \qquad \forall a \in RA \tag{4.2}$$

$$\sum_{s=0}^{L^a - 1} x_{t+s}^a \geq 1 \qquad \forall a \in RA, 1 \leq t \leq |T| - L^a + 1 \tag{4.3}$$

$$\sum_{t \in LC^a} z_t^a \geq 1 \qquad \forall a \in RA \tag{4.4}$$

$$z_t^a \leq x_t^a \qquad \forall a \in RA, t \in LC^a \tag{4.5}$$

$$x_t^m + x_t^n \leq 1 \qquad \forall t \in T, (m, n) \notin C \tag{4.6}$$

$$\sum_{t \in T_p} y_t^p = 1 \qquad \forall p \in PA \tag{4.7}$$

$$x_s^p \geq y_t^p \qquad \forall p \in PA, t \in T_p, s = t, ..., t + D_p - 1 \tag{4.8}$$

$$m_t \geq x_t^a \qquad \forall a \in A, t \in T \tag{4.9}$$

$$x_t^a, z_t^a, y_t^p, m_t \in \{0, 1\} \qquad \forall a \in A, p \in PA, t \in T, \tag{4.10}$$

The first two terms in the objective function are the sum of possession costs and the maintenance costs. In our formulation of the PMSP we require that the interval between successive executions of an activity is bounded, but it does not always need to be the same value. Accordingly, it is not clear beforehand how many executions will be in the planning horizon. The last term in the objective function is used to value the last interval. It is intended to eliminate the end-of-horizon effect, but it creates difficulty in modelling.

Constraints (4.2) guarantee that each work is carried out at least once in the possible truncated first planning cycle. Constraints (4.3) ensure, that the works until the end of planning horizon are scheduled at most L^a time periods from each other. Constraints (4.4)- (4.5) define the length of the last interval. Basically, if in the last planning cycle there are two or more executions of the same work, then z_t^a is set to one for only one time period t that results in the shortest remaining interval until the end of planning horizon. On the same link and at the same time only combinable activities can be carried out. This is ensured by constraints (4.6). These combinable jobs can be either routine works or projects. Constraints (4.7) guarantee that each project is executed once. Furthermore, constraints (4.8) ensure that each project is assigned to the right number of time periods and the starting time for performing the projects is in the interval (earliest possible starting time, latest possible starting time). Furthermore, these projects are assigned to subsequent intervals. Constraints (4.9) ensure that time period $t \in T$ will be occupied for preventive maintenance work if and only if for that time period on this segment at least one work is planned. Finally, constraints (4.10) ensure that the decision variables are binary.

If we consider an individual schedule of a given routine work, then the most cost effective way to schedule it is always at the maximum length of its planning cycle, that is, exactly at L^a periods. In this way, no extra maintenance work is done within the planning horizon, so the maintenance cost decreases, but there are less opportunities to combine executions of the works in one period, so the track possession time (or track possession cost) increases.

We extend the PMSP by restricting the time periods between two consecutive executions of the same work exactly to L^a time periods and we call this problem the *Restricted Preventive Maintenance Scheduling Problem (RPMSP)*, which is formulated as follows:

$$(RPMSP) \quad Min \sum_{t \in T} pc_t m_t + \sum_{a \in A} \sum_{t \in T} mc_a x_t^a + \sum_{a \in RA} \sum_{t \in LC^a} mc_a b_t^a x_t^a \tag{4.11}$$

s.t.

$$\sum_{t=1}^{L^a-G^a} x_t^a = 1 \qquad \forall a \in RA \tag{4.12}$$

$$x_t^a = x_{t+qL^a}^a \qquad \forall a \in RA, 1 \leq t \leq L^a, q \geq 1 \tag{4.13}$$

$$x_t^m + x_t^n \leq 1 \qquad \forall t \in T, (m,n) \notin C \tag{4.14}$$

$$\sum_{t \in T_p} y_t^p = 1 \qquad \forall p \in PA \tag{4.15}$$

$$x_s^p \geq y_t^p \quad \forall p \in PA, t \in T_p, s = t, ..., t + D_p - 1 \tag{4.16}$$

$$m_t \geq x_t^a \qquad \forall a \in A, t \in T \tag{4.17}$$

$$x_t^a, y_t^p, m_t \in \{0,1\} \qquad \forall a \in A, p \in PA, t \in T, \tag{4.18}$$

The objective again minimizes the sum of possession costs, the maintenance costs and the penalty cost paid if the last execution of the routine works is carried out too early in the planning horizon compared to the end of horizon. Constraints (4.12) ensure that each routine maintenance work is scheduled exactly once in the first allowed planning cycle and then constraints (4.13) guarantee that until the end of the planning horizon the works for the other cycles will be defined as well, ensuring exactly L^a time periods between two subsequent occurrences of the same job. Constraints (4.14)- (4.18) have been already explained before.

Theorem 4.1 *The Preventive Maintenance Scheduling Problem is NP-hard.*

Proof. It is well known that the following problem is NP-complete (see Garey and Johnson (1979)). Given an undirected graph G and a positive integer k, decide whether or not G can be coloured with k colours such that no edge is incident to vertices of the same colour. We show that it is NP-complete to decide whether the PMSP, as well as the RPMSP, has a solution of objective value 0 (*i.e.* a 0 cost solution).

Consider a graph $G = (V, E)$ and let $n = |V|$ be the number of vertices. We assume that $V = \{1, \ldots, n\}$. We construct an instance of the PMSP with $|T| = n$, with n routine works, each of them having $L^a = n$ and $G^a = 0$, and without projects. The i-th and j-th routine works have a conflict if ij is an edge in G. We assume that the maintenance costs are zero. The first k time periods have possession cost 0, the remaining $n - k$ periods have possession costs 1. This is an instance of both the PMSP and the RPMSP. The size of our construction is clearly polynomial in the size of graph G. We claim this instance has a solution with objective value zero if and only if graph G can be coloured with k colours. Indeed, suppose that G has a k-colouring. If a vertex has colour c, assign the corresponding routine work to the c-th period. Then we obtain an assignment without conflicts. This assignment has zero cost. Conversely, consider a zero-cost solution of the railway problem. Then each routine work is scheduled for any of the first k periods. Colour a vertex of G by colour c if its corresponding routine work is scheduled for the c-th period. This yields a k-colouring of graph G. □

4.5 Solution approach

The PMSP and RPMSP are modelled in GAMS and they are solved afterwards with the MIP solver CPlex 7.1. Solving these optimization problems, especially PMSP, to optimality for a single link, more than 15 types of maintenance works and for more than 3-4 years, requires a large amount of time. To improve the performance of the PMSP, we added to the model the following redundant constraints:

$$z_t^a + x_s^a \leq 1 \quad \forall a \in RA, t \in LC^a, t+1 \leq s \leq |T| \tag{4.19}$$

Constraints (4.19) guarantee that for each routine work $a \in RA$ if $d \in LC^a$ is the last time period when work a is carried out (*i.e.* $x_d^a = 1$), then $z_t^a = 0$, for all the time periods from the last planning cycle smaller than d. Constraints (4.19) improved somewhat the performance of the PMSP, but it is still impossible to get for some instances optimal solutions within 3 hours. More details about performances are presented in Section 4.7.

Since it takes too much time to find the optimal solution it might be better to settle for a non-optimal solution which has somewhat larger overall cost, but which is still quite close to the optimal objective value, and which can be found in a reasonable time. Therefore, our further purpose is to develop heuristics for solving the PMSP and RPMSP. In the literature we could not find any algorithms which can be used for solving our problems, since the problems for which those heuristics were developed are somewhat different from our problem.

In the following section we develop two heuristics for solving the PMSP and another two for solving the RPMSP. These heuristics are greedy in sense that they try to combine every activity together. In the next section each of them is presented in detail. It is worth mentioning that in each of the approximation methods we try to schedule the routine maintenance works and projects together. If two or more routine works cannot be combined then they will be scheduled for separate time periods.

4.6 Heuristics

First, we explain the two heuristics for solving the RPMSP, and thereafter the other two for the PMSP, since some steps in the first two heuristics are used in the last ones as well, namely, steps for making a preliminary schedule with restricted planning cycles.

Before we describe the heuristics, we recall the input of the problem and some notation. Given are a set of routine maintenance works RA with their planning cycle L^a and frequencies F^a, $\forall a \in RA$. In the beginning of the planning horizon T there are G^a periods elapsed since maintenance work $a \in RA$ was in the past for the last time carried out. A set of projects PA with duration D_p and a set of possible start points $T_p \in T$ are given. There is also a list of works which cannot be scheduled together. $e_{a,k}$ denotes the k-th execution time of work $a \in RA$, pc_t the possession cost in period $t \in T$ and mc_a cost for carrying out maintenance work $a \in A$.

The first heuristic, called Single Component Strategy (SCS) starts with making the best individual schedules for each of the routine works. In the literature this heuristic is sometimes called Decomposition approach. The idea of the SCS can be found also in inventory control as a strategy for independent ordering (see Chopra and Meindl (2001)). In the SCS the time periods between two consecutive executions of the same routine work are kept constant. The projects are added later to this schedule. In this algorithm we do not look at the possible combination of the works, we focus only on the individual plans. However, some works will be combined anyway due to the structure of the problem.

4.6.1 Single Component Strategy (SCS)

Step 0:
(For $j = 1, ..., n$) Make an individual schedule for work j such that the sum of the possession cost, maintenance cost and penalty paid for late execution per time horizon $|T|$ is minimized. These costs are calculated for each work separately, not looking at the savings in the possession cost resulting by combining some works. There might be certain periods where work j cannot be scheduled due to the earlier choices made for works $1, .., j-1$.
Step 1:
Choose a project p from set PA with the earliest possible starting time. In the allowed time interval T_p find the best time moment for performing it together, as much as possible, with already scheduled routine maintenance works.
Step 2:
Calculate the overall cost of the schedule resulting from Step 0 and Step 1.

The next heuristic, called Most Frequent Work First (MFWF), starts with scheduling the works having the highest frequency. The time periods between two consecutive executions of the same work are again kept constant. First, we make a schedule for the most frequent work. Then the other works are scheduled such that the increase in the overall cost is minimal. After the best possible schedule is found for the routine works, the projects are also scheduled, each of them in their allowed time intervals.

4.6.2 Most Frequent Work First heuristic (MFWF)

Step 0:
Order the set of routine works $RA = \{1, ..., n\}$ such that the planning cycles are in increasing order, *i.e.* $L^1 \leq L^2 \leq ... \leq L^n$. Schedule routine work 1 at its maximum interval, *i.e.* in periods $1, 1 + L^a, 1 + 2L^a, ...$ *etc.*
Step 1:
(For $j = 2, ..., n$) Schedule routine work j such that the increase in the sum of possession cost, maintenance cost and penalty cost for too late execution of work j in the planning horizon is minimal. The first execution time of work j should be in the period $[1, L^j - G^j]$ and there might be certain periods where work j cannot be scheduled due to the earlier choices made for works $1, .., j - 1$.
Step 2:
Choose a project p from set PA with the earliest possible starting time. In the allowed time interval T_p find the best time moment for performing it together, as much as possible, with already scheduled routine maintenance works.
Step 3:
Repeat Step 1 and 2 for all values where work 1 can be carried out for the first time, *i.e.* in the period $[1, L^1 - G^1]$. The schedule resulting in the minimum cost is chosen.

We expect that the second heuristic will perform better than the first one due to optimization done in Step 1 and Step 3. Our purpose with the first heuristic is to show how much the overall costs can be decreased if some global optimization steps are also included in the algorithm.

In the last two heuristics for solving PMSP, we allow shorter intervals between two consecutive executions of the same routine work, creating more possibilities to combine work. This results in lower possession cost, but in most of the cases higher maintenance cost. Both heuristics start with making a schedule with fixed intervals, using MFWF (SCS or the IP model for the RPMSP can be used as well). After that a better/cheaper schedule is searched by modifying, *i.e.* shortening, the planning cycles and creating more opportunities for combination of the works.

The idea of the third heuristic, which is called Opportunity Based heuristic (OBH) is coming from the Opportunity-based maintenance model (see *e.g.* Dekker *et al.* (1997)). In that model preventive maintenance is carried out at opportunities that are generated by failure of a particular unit in the system. Hence, planned maintenance activities are combined with unplanned activities. In the OBH, the execution times of the most frequent work will be used as opportunities for execution times for the other routine works as well. First of all, a preliminary plan is made, where we schedule first the works having the highest frequency. After that we check whether fitting all the execution times of the other works into the schedule of the most frequent work leads to a lower cost than creating separate, own opportunities and requiring more possessions.

4.6.3 Opportunity Based Heuristic (OBH)

Step 0:
Order the set of routine works $RA = \{1, ..., n\}$ such that the planning cycles are in increasing order, *i.e.* $L^1 \leq L^2 \leq ... \leq L^n$. Schedule these routine works together with the projects, using the MFWF heuristic. The execution times of the most frequent job give the initial values of the opportunities' list, $Opp = \{e_{1,1}, e_{1,2}, ..., e_{1,F^1}\}$.
Step 1:
(For $j = 2, ..., n$ and for all execution times of work j) If a given execution time of work j, $e_{j,k}$, is not yet in the Opp list, then check whether shifting forward $e_{j,k}$ to the closest earlier opportunity leads to a lower cost than the cost of a new possession. If $\frac{mc_j \cdot S}{L^j} > pc_t$, where S is the length of the time period between the current execution time and the closest earlier opportunity, then a new opportunity is created and $e_{j,k}$ is added to the Opp list. If not, then all the execution times of routine work j are shifted S time periods forward. In the latter situation sometimes forward shifts of the execution times of work j are not possible, because at that time the execution of another work, which cannot be combined with work j, has been already planned. If shifting the execution times forward is still possible, then adjust the frequencies as well, since shifting forward actual execution times might lead to more executions within the planning horizon. If in the new schedule of work j in the $L^j + 1$ consecutive time periods there are more than two works scheduled, then the middle execution can be always deleted from the schedule, decreasing the possession and maintenance cost. If the new schedule of work j results in a lower overall cost than before, then the schedule from Step 1 is modified with this new schedule, otherwise the schedule found in Step 1 is used.
Step 2:
Rescheduling of the projects is carried out according to Step 2 from the MFWF heuristic. The schedule resulting in the minimum cost is chosen.

The following heuristic (MCWF) is based again on the opportunity-based maintenance. In this case, the execution times of the most costly work and half of the time intervals between two subsequent executions are used as opportunities for execution times for the other routine works as well. Basically, MCWF starts with making a preliminary plan by scheduling the most costly works first. In the preliminary plan we shift the execution times of the other works such that in the end all works are carried out only in the listed opportunities, even if it results in a higher maintenance cost. The main difference between the last two heuristics is that in the first heuristic we shift the executions of the routine works to the closest earlier opportunity only if this is locally beneficial and in the second one we move always execution time forward, even if this action has a negative effect on the costs.

4.6.4 Most Costly Work First heuristic (MCWF)

Step 0:
Order the set of routine works $RA = \{1, ..., n\}$ such that their maintenance costs per planning horizon $|T|$ are in decreasing order, *i.e.* $mc_1F^1 \geq mc_2F^2 \geq ... \geq mc_nF^n$.
Step 1:
Schedule these routine works together with the projects, using the MFWF heuristic (with the RA set ordered in Step 0).
Step 2:
The execution times of the most costly job and the rounded down value of the average of its two consecutive execution times give the initial values of the opportunities list, $Opp = \{\lfloor t_{1,1}/2 \rfloor, t_{1,1}, \lfloor (t_{1,1} + t_{1,2})/2 \rfloor, t_{1,2}, ..., t_{1,F^1}\}$.
Step 3:
(For $j = 2, ..., n$ and for all execution times of work j) If a given execution time of work j, $t_{j,k}$, is not yet in the *Opp* list, then $t_{j,k}$ is shifted forward to the closest earlier opportunity even if this leads to a higher overall cost. Take into account that sometimes shifting forward once the execution time of work j is not possible, because at that time the execution of another work, which cannot be combined with work j, has been planned. If this is the case and there is still one more earlier opportunity, which does not coincide with $t_{j,k-1}$, then try to shift $t_{j,k}$ one more period forward. If this leads again to conflicting execution times, then we conclude that we cannot get a feasible solution. If shifting an execution time forward is still possible, then adjust the frequencies as well, since shifting forward actual execution times might lead to more executions. Check whether in the new schedule of work j in the $L^j + 1$ consecutive time periods there are more than two works scheduled. If yes, then delete the middle execution from the schedule, decreasing thereby the possession and maintenance cost.
Step 4:
Rescheduling of the projects is carried out according to Step 2 from the MFWF. The schedule resulting in the minimum cost is chosen.

As we already mentioned, all four heuristics are greedy heuristics, some of them containing improvement steps. The complexity of these algorithms is as follows: MFWF, OBH and MCWF have complexity of $O(n(n+p)T^3)$ and SCS $O(n(n+p)T^2)$, where n is the number of routine maintenance works, p is the number of projects and $|T|$ is the planning horizon.

	Scenario 1				*Scenario 2*			
	n=25		*n=15*		*n=25*		*n=15*	
	Average	*St.dev*	*Average*	*St.dev*	*Average*	*St.dev*	*Average*	*St.dev*
SFre	135	24.9	77.3	17.2	135	24.9	77.3	17.2
NV	2843.6	124.5	1768.2	97.4	5443.6	124.6	3328.2	97.5
NC	3860	203.6	2366.6	192	9866.9	3706.2	11315.8	2605.1

Table 4.1: Problem specifications. (*SFre*- the sum of works' frequencies, *NV*- the number of variables and *NC*- the number of constraints)

4.7 Computational results

Experimental set-up and implementation

The planning horizon for the generated instances is 2 years and the discrete time periods are weeks. Furthermore, we assume that each routine maintenance work has different planning cycles, and consequently, different frequencies and different maintenance costs. We assume that the track possession cost is the same for each week within the planning horizon.

To test the algorithms, we generated instances with $n = 15$ and $n = 25$ routine works. For each size we generated 10 instances. The generated values for the planning cycles (L^a), for the number of periods elapsed since maintenance work a was in the past for the last time carried out (G^a), for the maintenance costs (mc_a), for number of projects which have to be performed until the end of the planning horizon (p), the possible earliest (ES_p) and latest starting times (LS_p) and the duration of the projects (D_p) are uniformly distributed random numbers, as follows $L^a \sim U[4, 52]$, $G^a \sim U[0, L^a]$, $mc_a \sim U[1, 100]$, $p \sim U[0, 2]$, $ES_p \sim U[1, 104]$, $LS_p \sim U[1, 104]$, $D_p \sim U[1, 6]$.

We tested our models and algorithms for two scenarios. In the first scenario we assume that each routine work can be combined with all other routine works and projects, but the projects cannot be combined with other projects. In the second scenario we want to see the effect in the exact models and heuristics of having routine works that cannot be combined. Therefore, we assume in the second scenario that a group of two works and a group of three works cannot be combined. Thus the works within a group cannot be carried out at the same time. These works are arbitrary chosen. Furthermore, we test both models and all four heuristics for different possession costs ($PossC = 25$ and $PossC = 75$).

All the tests are executed on a Pentium IV 1.60GHz (256MB RAM) personal computer, using CPlex 7.1 for calculating the LP and the optimal integer solution. In Table 4.1 we show for both scenarios and possession costs the mean of the sum of the works frequencies (SFre), the average number of variables (NV) and the average number of constraints (NC). In scenario 2, where some routine works cannot be combined, the number of variables increased by almost 100%. The results, after running the model for the generated instances, are shown in Table 4.2 - 4.3.

Results for the RPMSP and PMSP

From the test results we conclude that increasing the possession cost decreases the number of possessions and increases the number of extra maintenance works. If the track possession is very cheap ($PossC < 10$) then the number of possessions and the number of activities are almost the same in both models. The PMSP results in 2% to 12% lower overall costs than the RPMSP, however the computational time is much longer and sometimes it is even impossible to get an optimal solution. Actually, RPMSP could be solved to optimality within 8 minutes for all the instances, even in the scenario where some works cannot be combined. However, in the case of the PMSP only in 29% of the instances we could get an optimal solution within 3 hours. Comparing Scenario 2 with 1 we can conclude that if some works cannot be combined, then the number of possessions, the overall maintenance cost and the relative difference (RdOLP) between the optimal value (VOpt) and the LP relaxation value (VLP) for both models increases with approximately $0.4\% - 2.4\%$. An important observation that was not reported in the tables is that the gap between VOpt and VLP is much bigger for the instances where besides the routine maintenance works there are some projects as well. In our opinion this is not a very shocking result, since from the beginning we knew that scheduling routine works and projects together gives a very complicated structure to the problem. On the other hand, combining routine works with projects means that once there is a possession for a given project, some other routine works can be carried out at the same time, saving some extra possessions.

Results for the heuristics

In Tables 4.2 - 4.3 the results for the heuristics are given. We recall that SCS and MFWF are heuristics developed for solving the RPMSP and OBH and MCWF are for solving the PMSP.

Table 4.2 shows that the MFWF heuristic performs well, for 15% of the instances we get the optimal solution, especially if the possession cost is relative small ($PossC < 15$). As the relative difference (RdOH) between the optimal value (VOpt) and the heuristic value (ObjV) shows, SCS performs quite poorly, it results in a 6% to 19% higher overall cost than the RPMSP. On average, savings from 5% to 15% can be achieved by using

Method *PossC*		*Scenario 1* *Average* n=25	n=15	*Scenario 2* *Average* n=25	n=15
RPMSP	VOpt	10030.9	5797.3	10077.3	5829.9
25	TCPU(s)	358.9	90.3	186.3	274.2
	VLP	9358.2	5295.9	9358.2	5295.9
	RdOLP(%)	6.7	8.6	7.1	9.1
SCS	ObjV	10645.3	6236.6	10775.5	6249.8
25	TCPU(s)	0.01	0.01	0.01	0.01
	RdOH(%)	6.1	7.6	6.9	7.2
MFWF	ObjV	10124.2	5848.9	10172.7	5886.1
25	TCPU(s)	0.013	0.014	0.013	0.012
	RdOH(%)	0.9	0.8	0.9	0.9
RPMSP	VOpt	12433.9	7667.3	12562.6	7759.8
75	TCPU(s)	442.2	81.9	150.4	269.1
	VLP	10921.7	6571.9	10921.7	6571.9
	RdOLP(%)	12.2	14.2	13.1	15.3
SCS	ObjV	14540.3	9131.6	14560.1	9169.8
75	TCPU(s)	0.01	0.01	0.01	0.01
	RdOH(%)	16.9	19.1	15.9	18.1
MFWF	ObjV	12669.9	7817.2	12832.5	7897.6
75	TCPU(s)	0.014	0.012	0.013	0.012
	RdOH(%)	1.9	1.9	2.1	1.7

Table 4.2: Computational results for RPMSP (*VOpt*- optimal value of RPMSP; *TCPU*- CPU time; *VLP*- LP relaxation value; *RdOLP*- relative difference between the VOpt and VLP; *ObjV*- solution value of the heuristic; *RdOH*- rel. diff. between VOpt and ObjV)

MFWF *versus* SCS. The CPU times (TCPU in seconds) for both heuristics are very small, 1-2 hundredths of seconds.

The solution is found right away also in case of OBH and MCWF heuristics. The RdOH and the relative differences (RdHLB) between the heuristic value and the best MIP lower bound found by CPlex after running PMSP for three hours show that OBH performs better than MCWF. Especially, if $PossC = 25$, then there is 0.5%-1.9% increase in the overall costs in comparison with VOpt for PMSP, *versus* 1.9%-6.3% increase in costs for MCWF. However, if the possession increases, then OBH does not perform as good as before, resulting in case of $PossC = 75$, an approximately 6% relative difference between the heuristic value and the best MIP lower bound found. The reason why this happens is that in Step 1 we decide whether it is locally worth to move forward an execution time or not, without checking what the consequences of this movement are for the whole schedule. Shifting forward an execution of a given work means that we might be forced to include another execution time to the end of planning horizon, increasing hereby the number of maintenance works, but reducing somewhat the length of the remaining interval from the last execution to the end of horizon. The increase in the overall costs is even more than

Method *PossC*		*Scenario 1* *Average* n=25	n=15	*Scenario 2* *Average* n=25	n=15
PMSP	VOpt	10036.3*	5681.4	10086.8*	5705.2
25	TCPU(s)	> 10800	9455.8	> 10800	8462.1
	VLP	9124.5	5116.6	9124.5	5116.8
	RdOLP(%)	9.1	9.9	9.5	10.3
	RdOLB(%)	4.2	2.9	4.2	1.7
OBH	ObjV	10085.2	5771.4	10132.9	5816.5
25	TCPU(s)	0.014	0.014	0.013	0.013
	RdOH(%)	0.5	1.5	0.5	1.9
	RdHLB(%)	4.9	4.7	4.9	3.7
MCWF	ObjV	10410.2	6041.3	10283.3	6030.3
25	TCPU(s)	0.015	0.011	0.016	0.01
	RdOH(%)	3.7	6.3	1.9	5.7
	RdHLB(%)	8.3	9.6	6.5	7.5
PMSP	VOpt	11574.7*	6720.7	11710.1*	6889.1
75	TCPU(s)	> 10800	7424.4	> 10800	7163.2
	VLP	10161.2	5974.2	10161.2	5974.2
	RdOLP(%)	12.2	11.1	13.2	13.2
	RdOLB(%)	5.9	2.6	5.3	2.3
OBH	ObjV	12365.9	7201.6	12505.6	7359.1
75	TCPU(s)	0.014	0.012	0.014	0.014
	RdOH(%)	6.8	7.1	6.8	6.8
	RdHLB(%)	13.5	10.1	12.8	9.3
MCWF	ObjV	12361.9	7524.3	12363.1	7811.3
75	TCPU(s)	0.012	0.012	0.012	0.015
	RdOH(%)	6.8	11.9	5.5	13.4
	RdHLB(%)	13.5	15.0	11.5	16.1

Table 4.3: Computational results for PMSP. ((*)- no optimal solution could be found within 3h; *VOpt*- optimal value of PMSP or the best value found within 3h; *TCPU*- CPU time; *VLP*- LP relaxation value; *RdOLP*- relative difference between the VOpt and VLP; *RdOLB*- rel. diff. between VOpt and the last MIP lower bound; *ObjV*- solution value of the heuristic; *RdOH*- rel. diff. between VOpt and ObjV; *RdHLB*- rel. diff. between ObjV and the last MIP lower bound)

6% for MCWF if the possession cost increases. In summary we can say that including some local improvements/optimization steps in the algorithm results in a lower overall cost (possession cost+maintenance cost+penalty paid for too early execution of the last work within the planning horizon).

4.8 Conclusions

Since rail is an important transportation mode, proper maintenance of the existing lines, repairs and replacements carried out in time are all important to ensure efficient operation. Moreover, since some failures might have a strong impact on the safety of the passengers, it is important to prevent these failures by carrying out in time and according to some predefined schedules preventive maintenance works. Since the infrastructure maintenance costs represent a large part of the total operating costs, there is a need for developing operations research tools, which help the maintenance planners to come up with optimal maintenance plans.

In this chapter we presented a mathematical programming formulation for the Preventive Maintenance Scheduling Problem. Maintenance works are assigned to different time periods (months/weeks), minimizing the track possession cost and the maintenance cost. Since the maintenance scheduling problem is a complex optimization problem and for a large set of instances it is difficult and time consuming to solve the problem to optimality, it was necessary to develop some approximation methods, which still give solutions close to the optimal ones. Two heuristics, namely MFWF and SCS were developed for RPMSP and other two, OBH and MCWF for the PMSP.

From the results we can conclude that MFWF gives the best results for solving the RPMSP and OBH for solving the PMSP. Furthermore, solving the PMSP for different instances, scenarios and different possession values results in, on the one hand much lower possession cost than RPMSP, but on the other hand higher maintenance cost. If we compare the average optimal value of the RPMSP and the best solution value found within three hours for the PMSP, then one can see that restricting the time periods between two consecutive executions of the same routine work constant until the end of horizon leads to an 1.8% to 12% increase in overall costs. In other words, having more freedom for choosing the execution times and increasing the possibilities to combine activities in one period results in a lower maintenance cost. However, if the number of routine works increases, then it is more and more difficult to solve the PMSP to optimality within three hours. If the planners want to have in very short time a good (close to optimal) schedule, then either RPMSP or one of the heuristics can be used effectively.

As a final remark we like to mention that the model presented in this chapter is just a basic model, but it can be extended to solve many types of practical problems, since in reality there are many more constraints that a maintenance planner has to take into account.

Chapter 5

Genetic and Memetic Algorithms for Scheduling Railway Maintenance Activities

Abstract

Nowadays railway companies are confronted with high infrastructure maintenance costs. Therefore good strategies are needed to carry out these maintenance activities in a most cost effective way. In this chapter we solve the Preventive Maintenance Scheduling Problem (PMSP) using genetic algorithms, memetic algorithms and a two-phase heuristic based on opportunities. The aim of the PMSP is to schedule the (short) routine activities and (long) unique projects for one link in the rail network for a certain planning period such that the overall cost is minimized. To reduce costs and inconvenience for the travellers and operators, these maintenance works are clustered as much as possible in the same time period. The performance of the algorithms presented in this chapter are compared with the performance of the methods from an earlier work, Budai *et al.* (2006), using some randomly generated instances.

This chapter is based on Budai, Dekker and Kaymak (2008).

5.1 Introduction

Nowadays railway companies are confronted with high infrastructure maintenance costs. For example, in the Australian freight operations the maintenance costs represent 25-35% of total train operating costs (Higgins (1998)) and in the Netherlands in 2003 the expenses of maintenance and renewal were approximately 295 million Euro for maintenance and

150 million Euro for renewal (Swier (2003)). Therefore good strategies are needed to carry out these maintenance activities in a most cost effective way. Besides the high maintenance costs, the other problem that the railway infrastructure manager is facing is that due to densely used railway network it is more and more difficult to find time periods for preventive infrastructure maintenance that are long enough. Moreover, these time periods should be chosen such that the train operation is not disturbed too much.

The IMPROVERAIL project(see Improverail (2002)) and Budai and Dekker (2002) show that preventive railway maintenance works are carried out in several countries during train service. In the actual train timetable possible possession allocations are scheduled for maintenance so that it should not affect regular train services too much. This can be done, however, for occasionally used tracks, which is the case in Australia and some European countries. Some important references in this respect are (Higgins (1998) and Cheung *et al.* (1999)). If tracks are used frequently, one has to perform maintenance during nights, when the train traffic is almost absent, or during weekends (with possible interruption of the train services), when there are fewer disturbances for the passengers (see *e.g.* Budai *et al.* (2006)). In that case one can either make a cyclic static schedule, which is made by Den Hertog *et al.* (2005) and Van Zante-de Fokkert *et al.* (2007) for the Dutch situation or a dynamic schedule with a rolling horizon, which has to be made regularly. The latter is described in Cheung *et al.* (1999).

The Preventive Maintenance Scheduling Problem in the literature

Initially, Budai *et al.* (2006) started to study the preventive maintenance scheduling problem (PMSP) for railway infrastructure, where a schedule for preventive maintenance activities has to be found for one link such that the sum of the track possession costs and maintenance costs is minimized. The possession costs are mainly determined by the possession time, which is the time that a track is required for maintenance and cannot be used for railway traffic. The preventive maintenance activities consist of small *routine works* and *projects*. The routine works (*e.g.* inspections or small repairs) are short, but frequent. They are scheduled from once per month to once in a year. The projects include longer renewal works (*e.g.* sleeper tamping, ballast cleaning), that are in general performed only once or twice every few years and they are triggered by condition measurements. Although the PMSP considers one rail link only, it can be extended to a network using the concept of the Single Track Grids (STGs) presented in Van Zante-de Fokkert *et al.* (2007).

Budai *et al.* (2006) show that the PMSP is an NP hard problem. Moreover, the authors provide a mathematical programming formulation for PMSP and using this formulation and the CPlex solver they attempt to find the optimal solution for their problem. In general the computation time to find the optimal solution is too high, therefore some greedy heuristics were developed. These heuristics give a feasible solution within a very short time. However, even the best heuristic sometimes differs up to 7% from the value of the best solution found using the exact method. Therefore, our objective is now to generate better results for the PMSP using other type of solution methods.

Several authors recommend using genetic (GA) and memetic (MA) algorithms for scheduling problems. Both algorithms are popular in maintenance applications because of their robust and fast search capabilities that help to reduce the computational complexity of large optimisation problems, such as large scale maintenance scheduling models. A detailed description of GA and MA is presented in Section 5.3.2 and Section 5.3.3. Here we explain briefly only some successful applications of GA and MA for maintenance scheduling problems, alike the PMSP.

Genetic algorithms in the maintenance scheduling literature

GAs, that is based on the principles of natural selection and genetics, were developed by Holland (1975) and are widely used since then by many researchers in different fields. Here we list some studies on maintenance optimization where GA was used with success.

Grimes (1995) deals with the problem of planning a specific track maintenance work, the track tamping, using a genetic algorithm. GA gives good results for short sections of track (10 miles) and poor results for long sections (50 miles). In Sriskandarajah *et al.* (1998) GA is used for scheduling the frequency-based overhaul maintenance of the rolling stock (*i.e.* trains). The computational results show that the GA gives close to optimal solutions for randomly generated problems with known optimal solutions. GA is used in Fwa *et al.* (1994), and later in Chan *et al.* (2001), for planning road-maintenance. Liu *et al.* (1997) and Morcous and Lounis (2005) present two different approaches for optimizing maintenance strategies of bridge deck networks. Moreover, GA proves to be a very powerful tool for maintenance scheduling in power systems too. Some important works in this respect are Negnevitsky and Kelareva (1999), Abdulwhab *et al.* (2004). In Munoz *et al.* (1997), Lapa *et al.* (2000) and Lapa *et al.* (2006) GA is used for maintenance scheduling in the nuclear power plan.

Memetic algorithms in the maintenance scheduling literature

In many articles in the literature the performance and effectiveness of GAs is often improved by incorporating a local search operator (e.g. tabu search, simulated annealing, hill climbing) into the GA by applying the operator to each member of the population after each generation. These approaches are called in the literature memetic algorithms (MAs). Here we highlight a couple of promising applications of the MAs in maintenance optimization.

Dahal and Chakpitak (2007) presents the application of the genetic algorithms, simulated annealing (SA) and their hybrid for generator maintenance scheduling in power systems. They show that the hybrid approach is less sensitive to variation of the GA and SA parameters and gives better averaged results than GA and SA. Burke *et al.* (1997a) deals with the thermal generator maintenance scheduling problem, where the maintenance of a number of thermal generator units is scheduled such that the maintenance cost is minimized and enough capacity is provided to meet the anticipated demand. This problem was earlier solved by traditional optimization techniques, such as integer programming, dynamic programming and branch and bound. For small problems these methods gave an optimal solution, but as the size of the problem increased, the size of the solution space increased exponentially and hence also the running time of these algorithms. To overcome this problem, in Burke *et al.* (1997a), simulated annealing, genetic algorithms, tabu search and a combination of simulated annealing and tabu search were implemented to solve the thermal generator maintenance scheduling problem. Their results show that the tabu search algorithm performs the best. The genetic algorithm was the worst performing algorithm for solutions with large numbers of feasible solutions, but performs slightly better than simulated annealing for problems with a small number of feasible solutions. Burke and Smith (1997b) solves the same problem by using memetic algorithm, i.e a genetic algorithm combined with tabu search. For small problems the memetic algorithm performs as good as simulated annealing and tabu search, but for large problems memetic algorithm outperforms both algorithms. In Burke and Smith (2000) the thermal generator maintenance scheduling problem is solved by using the memetic algorithm with three types of local search, namely simulated annealing, hill climbing and tabu search. The memetic algorithm with tabu search proves again to be the best method, followed by the memetic algorithm with hill climbing. The memetic algorithm with simulated annealing gives reasonable results, but the execution time is significantly higher than the execution times of the other two algorithms. The computational results are also promising when a memetic algorithm with tabu search, simulated annealing and hill climbing is used for solving maintenance scheduling problems for the National Grid in South Wales (see *e.g.* Burke and Smith (1999b)). Li *et al.* (2002) use a combination of genetic algorithm with tabu search for solving maintenance scheduling problem of oil storage tanks. It has been shown that this tabu-based genetic algorithm outperforms GA.

Due to the above described successful applications of the meta-heuristics on problems alike the PMSP, in this chapter we will focus on finding better results for PMSP using the GA and MA with three local search algorithms, namely the *Steepest Hill Climbing (SHC)*, *Simulated Annealing (SA)* and *Tabu Search (TS)*. Moreover, we develop a new method, called GA_OPP, where the preventive maintenance works are carried out at opportunities that are generated either randomly or by an already planned preventive maintenance work. GA_OPP is a two phase method. In the first phase opportunities are generated using GA and in the second phase all the executions of the preventive works are fitted as much as possible to these opportunities.

The idea of performing preventive works at opportunities has already been used in Budai *et al.* (2006) at the *Opportunity Based Heuristic (OBH).* In that paper the execution times of the most frequent routine work were used as opportunities for performing the other routine works as well. Savic *et al.* (1995a) and Savic *et al.* (1995b) formulate the opportunity based maintenance problem (OBM) as a set partitioning problem and solved it using GAs. The OBM is somewhat different than OBH, since in OBM the components in a system are divided into groups and as soon as a component of a group fails, all the components of the group are replaced. Hence, a failure of a component of a group is used as an opportunity to replace the rest of the components of this group. Thus, the problem is to find an optimal grouping of the components of a system such that the total maintenance cost is minimized. In Dekker and van Rijn (1996) a decision-support system (PROMPT) for opportunity-based preventive maintenance is discussed. PROMPT was developed to take care of the random occurrence of opportunities of restricted duration. Here, opportunities are not only failures of other components, but also preventive maintenance on (essential) components.

It is worth to mention that the PMSP is just a basic problem, but it can be easily extended to other types of practical problems, for instance scheduling the maintenance works while taking the available manpower into account.

The remainder of the chapter is organized as follows. In the next section we give a short problem description, followed by Section 5.3 where we discuss the solution approaches for PMSP, such as Opportunity Based heuristic, Most Costly Work First heuristic, Genetic algorithms, Memetic algorithms and the Two-phase opportunity–based heuristic. The computational results are presented and analyzed in Section 5.4. Finally, in Section 5.5 and we formulate our conclusions.

5.2 Description of the Preventive Maintenance Scheduling Problem (PMSP)

The aim of the PMSP is to give a schedule for preventive maintenance activities in a finite horizon, such that the maintenance works are clustered as much as possible in the same period and the overall cost is minimized. Clustering multiple maintenance activities in the same period leads to a reduction in the possession costs, since execution of a group of works requires only one track possession. Moreover, since maintenance works often require one or more set-up activities, such as crew and equipment travelling, carrying out multiple works simultaneously results in significant savings in the set-up costs.

The PMSP can be defined as follows. Consider a set of routine maintenance activities and projects over a finite planning horizon which is divided into a number of periods (*e.g.* weeks, months). For each routine work the *planning cycle* and the *age* is known. The planning cycle is equal to the maximum number of time periods between two consecutive executions. The age is defined as the number of time periods elapsed since the routine work was carried out for the last time. The duration, the earliest and latest possible starting times of each project are known as well. Moreover, some routine works and projects may be combined to reduce the track possession times, but others may exclude each other. We assume that a list of works (routine works and/or projects) that can be combined is given. The *maintenance costs* of each routine work and project and the costs of having a *track possession* in the planning period are known. Since our planning horizon is finite and the routine activities are repetitive, an end of horizon valuation is used. Thus, we assume that a *penalty* is paid if the last execution of the routine works is carried out too early compared to the end of planning horizon. Thus, the penalty cost for a given routine work is equal to its maintenance cost times the length of the remaining interval from the last execution until the end of planning horizon divided by the length of the planning cycle of this routine work. The goal of the PMSP is to schedule the given set of routine maintenance works and projects, such that the sum of track possession costs, maintenance costs and the penalties paid for too early executions is minimized.

The main problem with the PMSP is that we have a combination of repetitive work and once-only work. If we would have only repetitive work, then one could look for structure in the problem. An example could be power-of-two policies, where a base interval has been selected and every activity is executed at a power of two of this base interval. Roundy (1985) has shown that in case of inventory control the maximum loss of such a policy compared to the optimal policy is only 6%. However, the loss functions in case of inventory control are much flatter than in our case, so such a policy is not likely to perform well, especially if regular work also has to be combined with projects.

5.3 Solution approaches for PMSP

In this section we attempt to develop new techniques that generate better results for the PMSP than the methods used in Budai *et al.* (2006). First we recall briefly the already developed methods and after that we focus on GAs, MAs and the Two-phase opportunity–based heuristic.

5.3.1 Solution approaches in Budai *et al.* (2006)

Budai *et al.* (2006) present an exact method and two heuristics designed for the PMSP, namely the Opportunity Based heuristic (OBH) and Most Costly Work First heuristic

(MCWF). Both heuristics are greedy in the sense that they try to combine every maintenance work together.

In the OBH the execution times of the most frequent work create opportunities to carry out the other routine works as well. The most frequent work is carried out at exactly equivalent intervals and for all the other routine works it is checked whether it is cost effective to use these already created opportunities or rather creating own opportunities requiring more possessions. Finally, the projects are scheduled to the best place given the schedule of the routine works.

In the MCWF the execution times of the most costly work and half of the time intervals between two subsequent executions are used as opportunities for execution times of the other routine works. The restriction here is that each work can be carried out only at these opportunities, no matter whether it leads to a cost reduction or not. The projects are finally scheduled to the best place given the schedule of the routine works. A detailed description of these approaches can be found in the referenced article.

5.3.2 Genetic Algorithm (GA)

First we give a general description of the GAs, followed by a subsection where we describe the encoding of the chromosomes, the creation of the initial population, the fitness evaluation and the selection schemes used for the PMSP. Finally, we give a description of the implemented crossover and mutation operators.

5.3.2.1 General description of GA

Genetic algorithms use a direct analogy of natural behaviour. A solution of the problem is represented as a chromosome, using strings of binary digits. The process of translating a certain problem into chromosomes is called encoding. Some chromosomes are better than others, thus to test the performance of an individual chromosome, a fitness function is needed. The total number of chromosomes is called the population size. After an initial population is created, the crossover and mutation operators are used to get the next generations.

Crossover is the analogue of mating in nature. From the population two parents are chosen with a fitness dependent chance. Crossover of these two parents result into two children (offsprings). Because parents with good chromosomes are more likely to be selected for crossover, the average fitness of the children will generally be higher than that of the population of parents. Thus selecting the fittest parents, the good characteristics will survive over generations. This means that a number of iterations the population will converge to an optimal solution to the problem. To avoid that the process gets stuck in a local optimum, mutation of the chromosomes is used. This will guarantee diversity among the solutions and so genetic algorithms find global optima. To control the genetic process, several parameters need to be chosen. These are: the population size, number of iterations, the probabilities for the genetic operators, *etc.*

5.3.2.2 GA for PMSP

In the genetic algorithm used for the PMSP the chromosomes are represented as two dimensional binary arrays, where columns represent the weeks and rows represent the different routine works and projects. Thus an 1 at position (i, j) in the chromosome means that routine work i is planned for time period j.

In Figure 5.1 (figure originates from Negnevitsky (2005)), we present the steps that we follow in the GA used for the PMSP. These are actually the common steps used in the literature of GA (see *e.g.* Negnevitsky and Kelareva (1999), Sastry *et al.* (2005)). Each step is explained in details below.

Step 1: Create initial population

In *Step 1*, the initial population for the PMSP is created randomly in the following way. For routine works, the executions are determined by choosing at random a number between three-quarters of the planning cycle and the entire planning cycle. For the first execution from this selected number, the age will be subtracted in order to not violate the planning cycle restriction. If this number is negative, the first week is chosen. For projects, since the duration is fixed, choosing a starting period is sufficient. This period is chosen at random in the set of possible starting points, but taking into account that each project has to be finished before the end of the planning period.

Step 2: Fitness evaluation

In order to assign in *Step 2* a fitness to every chromosome, the total costs for the resulting solutions are computed. The total cost is the sum of the maintenance costs, possession costs and penalties resulting from the end-of-horizon validation. As we already mentioned in Section 6.2, some maintenance works cannot be performed together at the same time period. Unfortunately, it is not possible to be add this as a restriction to the GA. Therefore, in order to prevent solutions that violate such a combination we add to the overall costs a very high penalty cost each time that not combinable works are planned for the same time period. Since PMSP is a minimization problem and the GA maximizes fitness, the fitness is calculated as a reciprocal of the total cost.

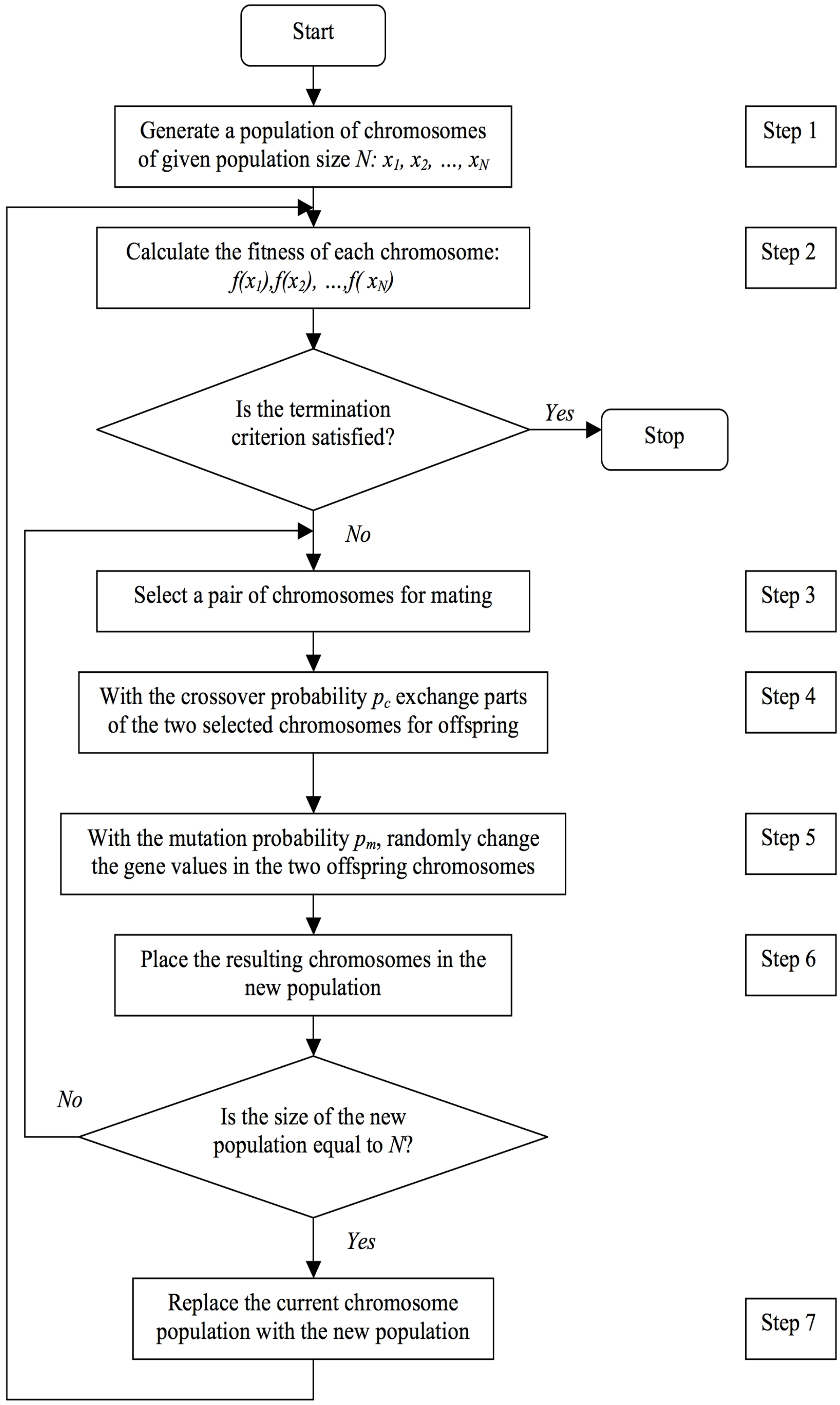

Figure 5.1: Schematic overview of a genetic algorithm (Negnevitsky (2005))

Step 3: Selection scheme

If the maximum number of iterations is reached, the algorithm stops, otherwise we proceed with *Step 3*, namely with selecting two chromosomes for mating. The selection of the chromosomes for producing offsprings is done by a selection scheme, such as the rank based roulette wheel selection, the tournament selection, *etc.* With the rank based roulette wheel selection, the chromosomes are ranked according to their fitness value, where the chromosome with the lowest fitness gets rank 1 and the one with the highest fitness gets a rank equal to the population size. The probability that a chromosome is selected is based on his rank. With the tournament selection, two chromosomes are selected at random. The one that has the highest fitness is selected as parent. After that two new chromosomes are selected at random in order to find the other parent.

Step 4: Crossover

One way to create offsprings from the chosen two chromosomes is by *one point crossover with indivisible parts*. Indivisible parts implies that the individual schedules of every routine work or project are kept equal throughout crossover, so that only the combination of the different routine works and projects is changed. Figure 5.2 gives an example of one-point crossover with indivisible parts.

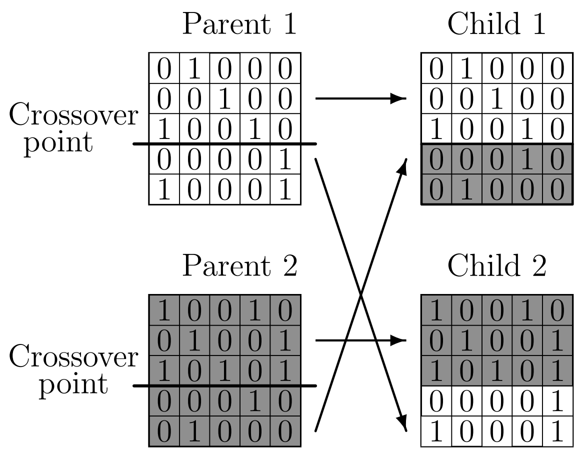

Figure 5.2: Crossover for GA

Step 5: Mutation

The other way to create offsprings is by using mutation operators. We recall, that the purpose of the mutation in the GA is to allow the algorithm to avoid local minima by preventing the population of chromosomes from becoming too similar to each other, thus slowing or even stopping evolution. We use five types of mutation operators. Three operators can be used for the routine works, one for projects and the fifth to the entire chromosome. These five mutation operators are described below.

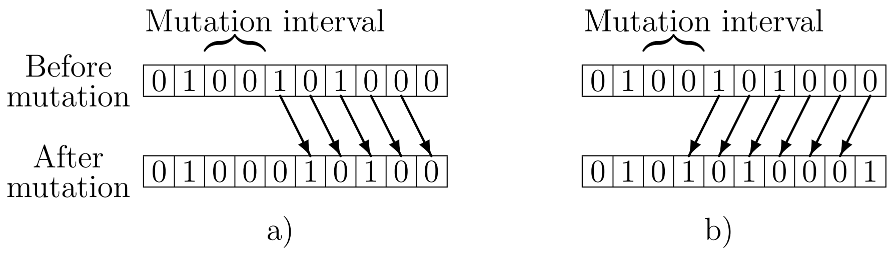

Figure 5.3: Zero insertion and zero removal mutation

The *zero insertion operator* (Figure 5.3a) tries to insert one or more zeros in one of the intervals between two executions of a routine work. First, it checks whether it is possible to insert a zero in any of the intervals, thus whether the second execution can be delayed one or more time periods. If this is not possible, there will be no changes. Otherwise, one of the intervals is selected at random and in this interval the operator checks how many zero's can be inserted in that interval at most. The number of zeros that will be inserted is a random number between one and the maximum number of zeros that can be inserted. After insertion all the elements after the inserted zero(s) are shifted to the right. If it is not possible to insert a zero in the random interval, a new random interval is chosen.

The *zero removal mutation operator* (Figure 5.3b) chooses an interval at random and in this interval a random number of zero's is removed. Thus, a number of executions of a routine work are carried out earlier than it was originally planned. If the last interval becomes too large due to the removal of some zero's, a new random execution is added.

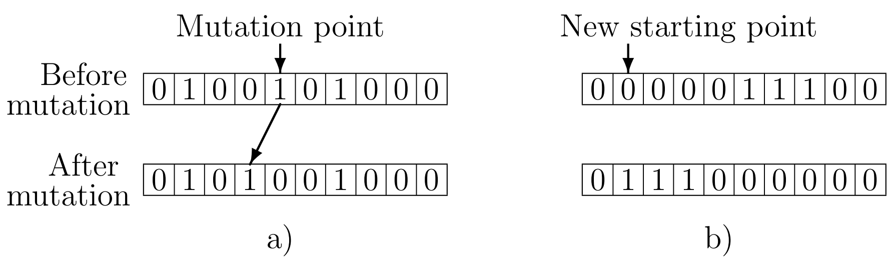

Figure 5.4: Single one shift and project mutation

The *single one shift mutation operator* (Figure 5.4a) chooses a random execution of a routine work and it checks whether it is possible to execute this routine work earlier or later. If this is possible then a random direction is chosen (forward or backward) and the execution is moved either forward or backward with a random number of places. All the other executions remain unchanged.

The three mutation operators described above only work on the routine maintenance works. They are actually shift operators, since the executions are shifted random time periods earlier or later. *Projects* have their own type of *mutation*, which chooses randomly a new starting time in the given (earliest allowed staring time, latest allowed staring time) interval. This is illustrated on Figure 5.4b.

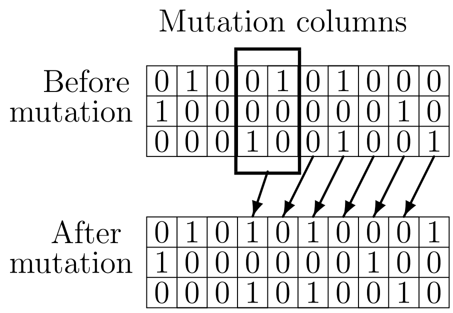

Figure 5.5: Join together columns mutation

The *join together columns mutation* (Figure 5.5) can be used for the whole chromosome, affecting all the routine works and projects. The following steps are used by this mutation operator:

a) Check whether there are two consecutive weeks with at least one execution of different works. If no, exit the algorithm. If yes, then one of these weeks is selected randomly and the corresponding columns are joined together in the chromosome.

b) Move all the executions after the merged columns one time period forward.

c) If the last interval for the routine works exceeds the cycle length, then the last column is filled with one, otherwise with zero.

d) Check if the starting times of all the projects are still in their allowed intervals. If not, reschedule the projects as it is described at the project mutation.

The crossover and mutation operators are performed with user defined probabilities. After selecting two parents out of the current population, it is first determined whether crossover will be performed on these parents (Step 4). If so, these two parents are replaced by their two children. After that, either both children or both parents (if the crossover was not selected) will be mutated individually (Step 5). The join columns together mutation is a global operator, so it cannot be used together with the other mutation operators, while the other four operators can be performed simultaneously on the same chromosome. Therefore, based on the user defined probability it is first determined whether the join columns together mutation will be performed on the chromosome. If not, it is decided randomly on which of the routine works which of the three mutation operators will be performed or which projects will be affected by the project mutation. If a mutation took place, then the original chromosome is replaced with the mutated one. Finally, the mutated chromosome is added to the new population.

Step 6: Add chromosomes to the new population

In order to improve efficiency and converge speed we implement elitism as well. The main idea of elitism is to preserve the best genetic material by copying the best member of each generation into subsequent generations. Thus, before any new chromosomes are added to the new population, a fixed number of chromosomes with the highest fitness are copied from the old population to the new population. The Steps 3 - 6 are repeated until the new chromosome population size is equal to the size of the initial population.

Step 7: Replace the old population with the new population

Finally, in Step 7 the current chromosome population is replaced with the new population and the Steps 2 - 7 are repeated until the maximum number of iterations is reached.

5.3.3 Memetic Algorithm (MA)

The memetic algorithm, as a variant of the genetic algorithm, is called a hybrid evolutionary technique (Burke and Smith (2000)), where the hybridization is realized by integrating the genetic algorithm with local search techniques. The memetic algorithm modifies the genes in order to get an offspring with a higher fitness. The modification of genes is accomplished by using local search on every produced new chromosome before adding it to the new population, so after Step 2 and Step 5 in Figure 5.1.

In this chapter we implement three local search algorithms, namely the *Steepest Hill Climbing (SHC)*, *Simulated Annealing (SA)* and *Tabu Search (TS)*. These are the most common local search techniques in the literature and Burke and Smith (1997b), Burke and Smith (1999b) have used these kind of algorithms with success on problems alike the PMSP. Burke and Smith (1997b), Burke and Smith (2000) affirm that the memetic algorithm is less sensitive to the quality and diversity of the initial population than the genetic algorithm. In spite of this observation, we still perform a local search on the randomly created initial population before starting the evolutionary part of the algorithm. In the literature, Burke and Smith (1997b), Burke and Smith (2000), Burke and Smith (1999a) and Burke and Smith (1999b) do not apply this initial local search, but Radcliffe and Surry (1994) do add this step.

As in Burke and Smith (2000), Burke and Smith (1999a) and Burke and Smith (1999b), we also implement here an iterative heuristic (IH) using the three local search heuristics mentioned above. The iterative heuristic repeatedly applies the local search optimiser to randomly generated solutions for the same number of times that the equivalent memetic algorithm applies local search. Thus, IH gives us possibility to compare the memetic algorithm with the repeated application of the individual local search operators.

All local search methods start looking for better solutions in the so-called *neighborhood* of the current solution. The neighborhood of a solution is the set of solutions that can be obtained by applying a very small local change to it. We implement the neighborhood in the following way. For routine works each execution is first moved one position to the left and thereafter one position to the right, unless the stopping criteria is met. We have three stopping criteria's, namely if a violation in the planning cycle occurs or the end of the planning horizon is reached or another execution is encountered. For projects, the starting time is set to one time period earlier or later, unless this new starting time is not in the given (earliest allowed staring time, latest allowed staring time) interval. Each change in the chromosome implies a new neighbor. Because of performance reasons instead of saving the exact neighbor we save just the move needed to create it.

As mentioned before, in this chapter we focus on three local search algorithms, namely on Steepest Hill Climbing (SHC), Simulated Annealing (SA) and Tabu Search (TS). These methods are described below.

5.3.3.1 Steepest Hill Climbing

The steepest hill climbing algorithm for the PMSP is comparable with the algorithm used in Moscato and Schaerf (1997). The following steps are performed:

1. Set the current solution.
2. Compute the difference in costs for all members of the neighborhood and choose the one that implies the highest cost reduction. If there are more than one solutions with the highest reduction, then one of them is chosen randomly.
3. If the best neighbor is better than the current solution then return to step 1. If the original solution is better than all its neighbors, the local optimum is found and the algorithm terminates as well.

5.3.3.2 Simulated Annealing

The simulated annealing algorithm used in this chapter is identical with the one from Moscato and Schaerf (1997) and Wolsey (1998). The main steps of the algorithm are described below.

1. Set the current solution.
2. Check whether the current temperature T is higher than the temperature limit. If not, then stop.
3. Given the current solution, choose randomly one of its neighbors.
4. Compute the difference in costs between the current solution and the selected neighbor. If the neighbor is better than the current solution, then the neighbor is accepted. Otherwise, the neighbor is accepted with a probability equal to $e^{-\Delta/T}$, where Δ is the difference in costs.
5. If the number of iterations for the current temperature has been reached, then the current temperature is decreased by the cooling rate α. Thus, $T_n = T_{n-1} \cdot \alpha$, where $0.8 < \alpha < 1$. Return to step 2.

5.3.3.3 Tabu Search

The basic principle of tabu search is to pursue the search whenever a local optimum is encountered by allowing non-improving moves. Cycling back to previously visited solutions is prevented by a tabu list, that records the recent history of the search. There are three parameters used for the tabu search, namely the tabu list length, the number of neighbors at each iteration and the total number of iterations.

The tabu search algorithm implemented for the PMSP can be described as follows:

1. Given the current solution, initialize an empty tabu list of a given length
2. Test whether the total number of neighbors is not smaller than the given number of neighbors to visit. If so, set the number of neighbors to visit equal to the total number of neighbors for this iteration of the algorithm.
3. For the number of neighbors to visit:
   a) Pick a random neighbor that has not been visited yet.
   b) Set the status of the picked neighbor to visited.
   c) Compute the costs difference for this neighbor.
   d) If this neighbor does not imply a cost reduction, check its tabu status.

e) If this neighbor is not tabu or does improve the original solution and it is the best neighbor found so far, set this as the current best neighbor.

4. Update the tabu list. If the list is full, replace the oldest entry by this new one.
5. Create a neighborhood for the picked solution and return to step 1, unless the given maximum number of iterations has been reached.

For a complete detailed description of the tabu search algorithm we refer to Moscato and Schaerf (1997) and De Ree (2006).

5.3.4 Two-phase opportunity–based heuristic (GA_OPP)

Since in Budai *et al.* (2006) the Opportunity Based Heuristic (OBH) gave in general good results for solving PMSP, we think that it is worth to use the idea of performing preventive maintenance works at opportunities in combination with the genetic algorithm. Therefore, in this chapter a third heuristic approach, called GA_OPP, has been developed to solve the PMSP. GA_OPP is a two phase method that is based on opportunities. In the first phase, opportunities are created using the genetic algorithms and in the second phase all the executions of the preventive works are fitted as much as possible to these opportunities.

In the GA_OPP the chromosomes are represented as one dimensional binary arrays, where the columns represent the time periods. Thus, the j-th element indicates whether at period j there is an opportunity. Because the chromosomes contain all the opportunities, we will also call them as opportunity lists. To each chromosome a two dimensional array is assigned, which is used once the maintenance activities are fit to the opportunity lists. In this array the resulting maintenance planning is created, so an element on position (i, j) equal to 1 means that routine work i is planned to be executed in time period j.

The GA_OPP developed in this chapter follows the steps presented in Figure 5.1. Some of the steps are exactly the same as in Section 5.3.2 (*e.g.* selection scheme), but others are completely different (*e.g.* generation of the initial population, crossover and mutation operators). Each step used for GA_OPP is explained in details below.

Chromosome/Opportunity list

0	1	0	0	1	0	1	0	0	1

Maintenance schedule

										(PlanCycle,Age)
0	1	0	0	1	0	1	0	0	1	(3,1)
0	1	0	0	1	0	1	0	0	0	(4,2)
0	1	0	0	0	0	1	0	0	0	(5,2)
0	1	0	0	0	0	1	0	0	0	(7,4)

Figure 5.6: Example: chromosome and the attached maintenance schedule for GA_OPP

Step 1: Create initial population

The initial population is created randomly, but with the following restrictions:

1. Check beforehand the minimum time between two consecutive executions of the preventive routine maintenance activities (T_{min}), *i.e.* the smallest planning cycle.

2. Calculate the minimum of the difference between planning cycle and age over all activities. This is denoted by ${T^1}_{min}$.

3. Order the routine maintenance works such that their planning cycles are in increasing order.

4. Until the population size is reached, do the following steps:

   a) Draw $\lceil \frac{|T|}{T_{min}} \rceil$ random numbers in the $[1, |T|]$ interval, where $|T|$ is the planning horizon. A random number j means that at time period j there is an opportunity, so the j-th element of the chromosome becomes 1.

   b) Check whether the gap between the first opportunity and the beginning of the time period is smaller than ${T^1}_{min}$. If not, then at time period ${T^1}_{min} - 1$ a new opportunity is created.

   c) Check whether all the gaps between two consecutive opportunities of the chromosome are smaller than T_{min}. If not, then for each gap larger than T_{min} a new opportunity is repeatedly created at $T_{min} - 1$ periods from the beginning of the gap.

5. Until the population size is reached, the preventive maintenance activities are fit to each chromosome in the following way:

   a) for each routine maintenance activity and according to its planning cycle choose the last allowed opportunity in the opportunity list.

   b) plan each project based on its own duration and starting time, such that the already existing opportunities are used as much as possible. If for fitting the projects new opportunities are needed, then these new opportunities are added to the opportunity list.

As the example in Figure 5.6 shows, in the maintenance schedule only the opportunities from the chromosome are used as possessions.

Step 2: Fitness evaluation

For GA_OPP the fitness of a chromosome is equal to the fitness of the attached maintenance schedule and that is calculated in the same way as we have already described in Section 5.3.2 for GA.

Step 3: Selection scheme

The selection schemes for the GA_OPP are identical with those from Section 5.3.2.

Step 4: Crossover

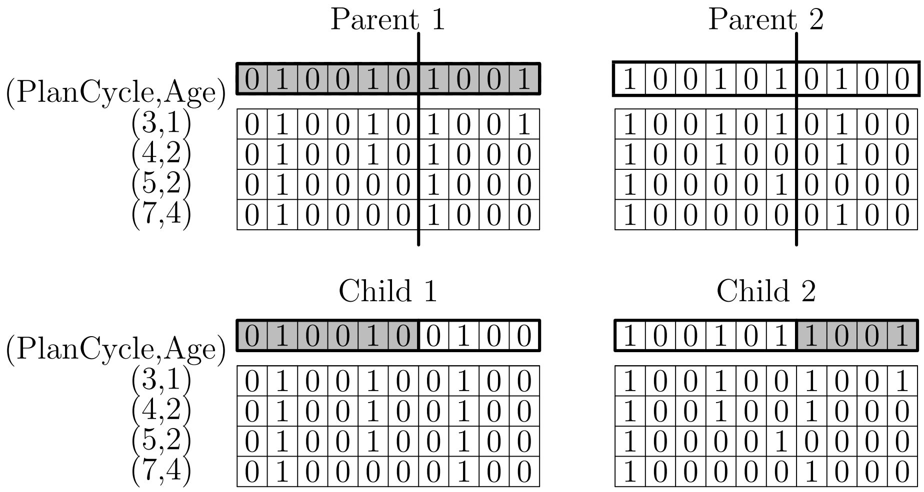

Figure 5.7: Crossover for GA_OPP

The *one point crossover* operator (see Figure 5.7) starts by choosing a random number between 1 and the length of planning horizon. Based on this crossover point the two chosen chromosomes are divided in two parts. For creating the two offsprings we interchange the opportunity lists of the two parents before and after this random crossover point. This might result in not feasible offsprings, since gaps between two consecutive opportunities might be longer than T_{min} or the gap between the first opportunity and the beginning of the time period is smaller than $T^1{}_{min}$. If such gaps are detected, then new opportunities are added to the opportunity list based on the Steps 4b and 4c used in creation of the initial population for GA_OPP. Finally, the maintenance works are fitted to the newly created two offsprings.

Step 5: Mutation

Beside crossover, mutation of the chromosomes is another way to create offsprings. For GA_OPP we use four types of mutation operators, namely *remove excess opportunities*, *insert opportunities*, *postpone executions* and *remove excess executions*. The first mutation operator might change the chromosomes, but not the maintenance schedule attached to the chromosomes, while the last three operators might change the maintenance schedule too. Actually, the last two operators make first possible changes in the maintenance planning attached to the chromosome and after that the chromosome will be updated, too. Below we describe each of these operators in details.

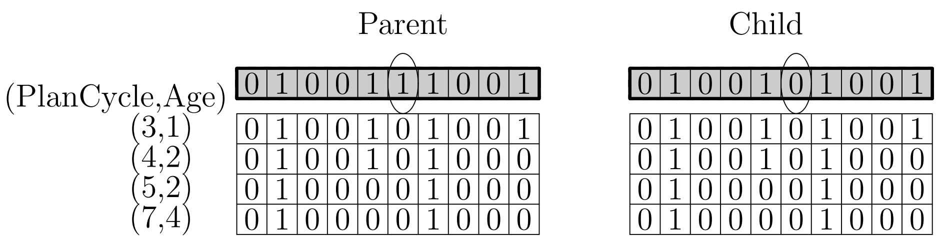

Figure 5.8: Remove excess opportunities mutation

The *remove excess opportunities operator* (Figure 5.8) deletes all the opportunities from the chromosome that have not been used for the maintenance planning.

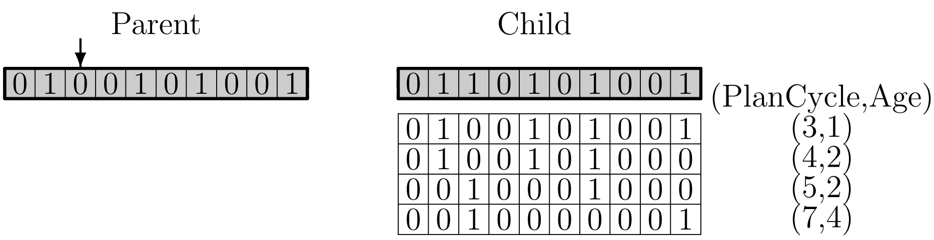

Figure 5.9: Insert opportunities mutation

With the *insert opportunities operator* (Figure 5.9) new opportunities are added to the chromosome. For each element of the chromosome and based on a user defined probability value it is decided whether that element will be changed into a new opportunity (unless it is already an opportunity) or not. Thereafter, the maintenance activities are fit to the chromosome.

The *postpone executions operator* (Figure 5.10) checks first in the maintenance schedule whether there exist possessions for which only one maintenance work is planned. If

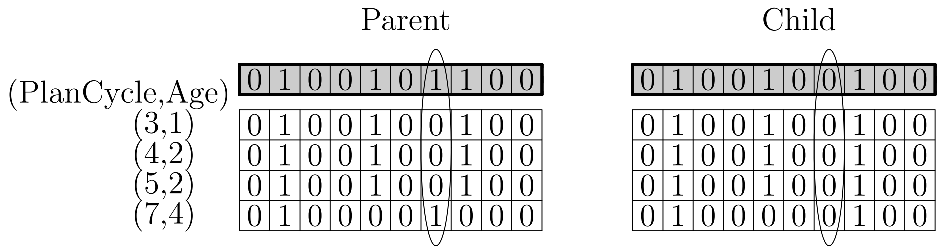

Figure 5.10: Postpone executions mutation

there are such of possessions, then it is tried for each of these possessions to postpone the single executions to a next opportunity, unless the stopping criteria is met. If executions could be postponed, then this might result in unused opportunities. Therefore, the chromosome will be updated as well.

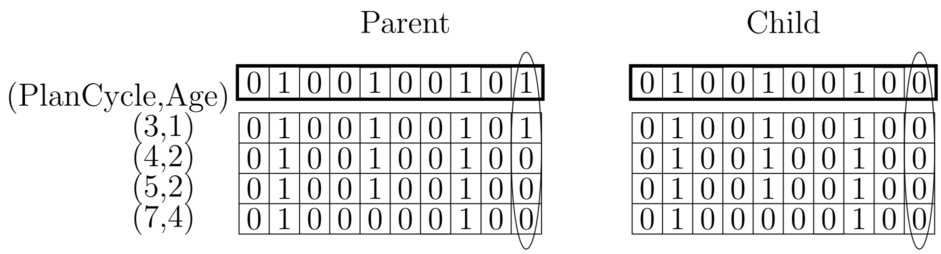

Figure 5.11: Remove excess executions mutation

Finally, the *remove excess executions operator* (Figure 5.11) checks in the maintenance schedule for each routine work whether there are excess executions and if so, then these executions are deleted. A given execution of a routine work is called excess execution, and thus it can be removed, if the time period between the next execution and the previous execution of the same work is smaller then the planning cycle. Removing executions might result in useless opportunities, therefore in the end the chromosome will be updated, too.

The crossover and four mutation operators are performed with user defined probabilities. First two parents are selected from the current population and after that it is determined whether crossover will be performed on these parents. If so, these two parents are replaced by their two children. After that, either both children or both parents (if the crossover was not selected) will be mutated individually. Based on the user defined probability value it is determined which of the four mutation operators will be used to generate new offsprings. If a mutation took place, then the original chromosome (and the original maintenance schedule) is replaced with the mutated one. Finally, the mutated chromosome is added to the new population.

Step 6: Add chromosomes to the new population

For GA_OPP heuristic the elitism is implemented, too. This step is identical with Step 6 described in Section 5.3.2 for GA.

Step 7: Replace the old population with the new population

In Step 7 the current chromosome population is replaced with the new population and Steps 2 - 7 are repeated until the maximum number of iterations is reached.

5.4 Computational results

Data

In this section we present the results of solving PMSP using the genetic algorithm, memetic algorithm, iterative heuristic and the two-phase opportunity–based heuristic. To be able to compare these results with the solution found in Budai *et al.* (2006) by the exact model and the greedy heuristics, we use the same problem instances and scenarios as in the referenced article.

We recall, that these are twenty randomly created problem instances, ten instances with 15 and ten instances with 25 maintenance works. The models and heuristics are tested for two scenarios and for two different possession costs ($PossC = 25$ and $PossC = 75$). In the first scenario each routine work can be combined with all other routine works and projects, but projects cannot be combined with other projects. In the second scenario it is assumed that for one particular group of two works and one particular group of three works, the works within these groups cannot be carried out at the same time. The works within such a group are randomly chosen.

The planning horizon for the generated instances is 2 years and the discrete time periods are weeks. We assume that the track possession cost is the same for each week within the planning horizon. Furthermore, we assume that each routine maintenance work has different planning cycles, ages, maintenance costs and each project has a different duration, possible earliest and latest starting time. All the tests are executed on a Pentium IV 2.5GHz 261MB RAM.

Parameter settings

In GA, MA, IH and GA_OPP there are a couple of parameters that need to be set by the user. These are not only the probability values for crossover or mutation, but also parameters as population size, elite size, number of iterations, *etc.*

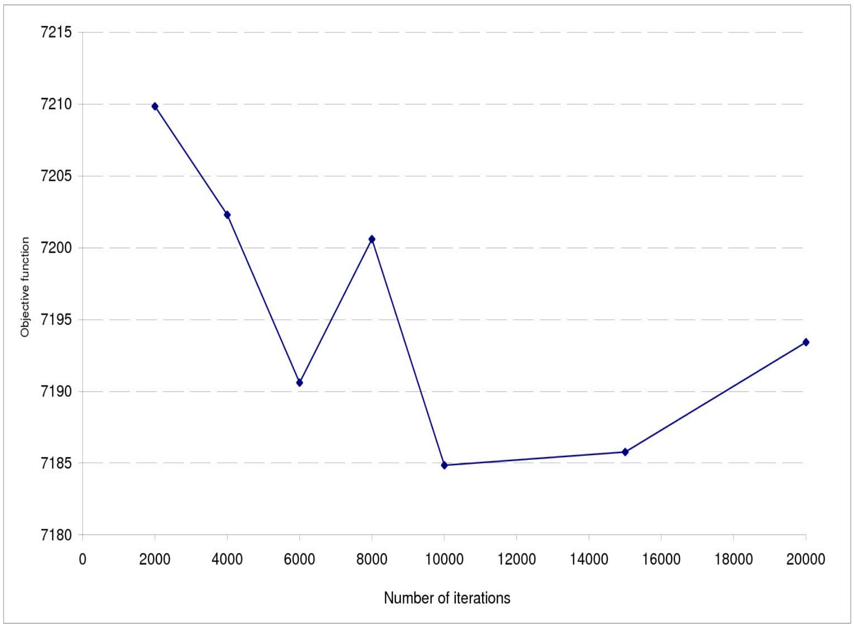

Figure 5.12: Results of GA for different number of iterations

The parameters used for GA, MA, IH and GA_OPP are determined after several experiments. One of the experiments that we carried out is to find out the right number of iterations for the GA. Figure 5.12 shows the results of different runs with different number of iterations. It turns out that number of iterations equal to 10000 results in the lowest objective value. In order to be able to compare the results for GA with the GA_OPP we decided to use the same number of iterations for GA_OPP, too. Moreover, for the GA and GA_OPP the population size and the elite size have been chosen to 100 and 2, respectively, based on a range of experiments. Furthermore, the rank based roulette wheel proved to be the best selection method.

The probability of crossover has been chosen for GA and GA_OPP equal to 0.8. For GA the probability of zero insertion, zero removal, project mutation and join together columns

Parameters	GA	GA_OPP	MA/IH SHC	MA/IH SA	MA/IH TS
Nr. iterations	10000	10000	2000	2000	2000
Population size	100	100	100	100	100
Elite size	2	2	2	2	2
Selection scheme type	rank based roulette wheel	rank based roulette wheel			
Prob. crossover	0.8	0.8	0.8	0.8	0.8
Prob. zero insertion	0.001		0.001	0.001	0.001
Prob. zero removal	0.001		0.001	0.001	0.001
Prob. shift single one	0.05		0.05	0.05	0.05
Prob. project mutation	0.001		0.001	0.001	0.001
Prob. join together columns	0.001		0.001	0.001	0.001
Prob. opportunity insertion		0.01			
Prob. remove unused opportunities		0.8			
Prob. postpone executions		0.8			
Prob. remove excess executions		0.8			
Starting temperature				2	
Cooling rate				0.9	
Nr. iterations per temperature				10	
Temperature limit				0.4	
Tabu list length					10
Size of subset of neighbors					15
Tabu search nr. of iterations					15

Table 5.1: Parameters used for GA, GA_OPP, MA and IH (SHC, Steepest Hill Climbing; SA, Simulated Annealing; TS, Tabu Search)

mutation operators are set to 0.001 and the probability of shifting a single execution to 0.05. The probability of mutation for GA_OPP are chosen slightly higher than for GA, namely the probability of removing unused opportunities, postponing executions and removing excess executions are set to 0.8 and the probability of inserting opportunities to 0.01.

The number of iterations for the memetic algorithm using the three types of local search optimizers is chosen equal to 2000. For the tabu search the length of the tabu list is set to 10, the number of iterations and the size of subset of neighbors for the tabu search is set to 15. Simulated annealing has only one tunable parameter, which is the cooling rate. Based on different empirical tests the cooling rate is fixed to 0.9. The initial temperature is set to 2 and as the algorithm progresses this temperature reduces gradually. In total 15 temperature levels are visited, resulting in a temperature limit of 0.4. For each temperature 10 runs are performed. All these parameter values were identified to be reasonable for the test problems.

All the parameters used for GA, GA_OPP, MA and IH are summarized in Table 5.1. These parameters are found to be those that gave the best results for GA, GA_OPP, MA and IH.

Method	Overall best	Best heuristic
CPlex	22 (14 opt)	-
OBH	0	1
MCWF	0	1
GA	0	0
GA_OPP	40 (3 opt)	51 (3 opt)
MA(TS)	15	16
MA(SA)	7 (2 opt)	10 (2 opt)
MA(SHC)	3 (2 opt)	3 (2 opt)
IH(TS)	0	0
IH(SA)	0	0
IH(SHC)	0	0

Table 5.2: Number of times a method proofs to be the best overall method or the best heuristic

Analysis of the results

The exact method (PMSP) and the two greedy heuristics (OBH - Opportunity Based Heuristic and MCWF - Most Costly Work First) from Budai *et al.* (2006) are compared here with the 8 new algorithms, namely with the genetic algorithm (GA), with the memetic algorithm using three local search algorithms: tabu search (MA(TS)), simulated annealing (MA(SA)) and steepest hill climbing (MA(SHC)), with the iterative heuristic using tabu search (IH(TS)), simulated annealing (IH(SA)) and steepest hill climbing (IH(SHC)) and finally with the two-phase opportunity–based heuristic (GA_OPP). These 8 algorithms are tested on all the 80 problem instances. The best solution for a given problem instance is chosen from the solutions of 10 runs of the algorithm on this problem instance.

Table 5.2 shows that the CPlex solver returns a solution that is not improved by any other method in 22 of the 80 cases. However, for 3 instances the GA_OPP and for 2 instances both MA(SA) and MA(SHC) find the same (optimal) solutions as the CPlex solver finds. We recall that the exact method (CPlex) could find only in 14 out of 80 instances an optimal solution within 3 hours. Since the CPlex solver was stopped after 3h, the CPlex solutions might be improved by any of these new heuristics, which is actually the case for 72.5% of the instances.

If we exclude the exact model from the comparison then for 63% of the instances GA_OPP gives the best solution, outperforming far the rest of the solution methods. Moreover, integrating the genetic algorithm with local search techniques gives better results than just simply performing genetic algorithm. As Table 5.2 shows, in any of the 80 cases the genetic algorithm and the three iterative methods do not return a best solution. The three memetic algorithms return however good results, the memetic algorithm with the tabu search outperforming the MA(SA) and MA(SHC), as Burke and Smith (1997b) and Burke and Smith (2000) conclude too. However, the MA(SA) and MA(SHC) find for 2 instances the optimal solutions.

	Method	Method performs worse										
		CPlex	OBH	MCWF	GA	GA_OPP	MA (TS)	MA (SA)	MA (SHC)	IH (TS)	IH (SA)	IH (SHC)
Method performs better	CPlex	-	70	79	54	22	39	35	39	80	80	80
	OBH	10	-	70	11	2	3	4	5	80	80	80
	MCWF	1	10	-	3	1	3	3	3	58	57	58
	GA	26	69	77	-	5	1	1	7	80	80	80
	GA_OPP	58	78	79	75	-	59	61	68	80	80	80
	MA(TS)	41	77	77	79	21	-	31	45	80	80	80
	MA(SA)	45	76	77	79	19	49	-	56	80	80	80
	MA(SHC)	41	45	77	73	12	35	24	-	80	80	80
	IH(TS)	0	0	22	0	0	0	0	0	-	40	45
	IH(SA)	0	0	23	0	0	0	0	0	40	-	43
	IH(SHC)	0	0	22	0	0	0	0	0	35	37	-

Table 5.3: Comparison of pairs of methods used for solving PMSP

In Table 5.3 we compare pairs of methods used for solving PMSP, because this provides good insight into the quality of the algorithms. It seems that GA performs better than OBH and MCWF, for 26 instances it gives even better results than the exact method. Furthermore, GA outperforms GA_OPP only for 5 instances and for each of the 80 instances the iterative heuristic. Actually, all the three iterative heuristics perform very poor, IH(SHC) being the worst out of these three methods. Comparing in Table 5.3 the performance of MA(TS), MA(SA) and MA(SHC), we can conclude that MA(SA) is equally good as MA(TS), but slightly better than MA(SHC).

In Table 5.4 and Table 5.5 the average results for all methods is given for each scenario separately, for different number of routine works and possession costs. Table 5.4 summarizes the results for the scenario 1 and Table 5.5 for scenario 2. In these tables we present

Method	PossC	SCENARIO 1					
		n = 25			n = 15		
		VOpt/ObjV	TCPU(s)	RdOH(%)	VOpt/ObjV	TCPU(s)	RdOH(%)
CPlex	25	10036.37*	>10800	-	5681.4	9455.84	-
OBH	25	10085.22	1.4	0.49	5771.43	1.4	1.58
MCWF	25	10410.24	1.5	3.73	6041.28	1.1	6.33
GA	25	9977.44	1020	-0.59	5694.52	1020	0.23
GA_OPP	25	9858.61	2760	**-1.77**	5611.87	1800	**-1.22**
MA(TS)	25	9893.81	1800	-1.42	5643.36	1200	-0.67
MA(SA)	25	9890.24	1980	-1.46	5640.46	1260	-0.72
MA(SHC)	25	9913.29	1560	-1.23	5645.29	780	-0.64
IH(TS)	25	11027.73	1380	9.88	6332.83	900	11.47
IH(SA)	25	11032.09	2580	9.92	6336.83	1440	11.54
IH(SHC)	25	11033.63	4740	9.94	6334.97	1680	11.50
CPlex	75	11574.71*	>10800	-	6720.74	7424.45	-
OBH	75	12365.91	1.4	6.84	7201.67	1.2	7.16
MCWF	75	12361.87	1.2	6.8	7524.28	1.2	11.96
GA	75	11947.99	1500	3.22	7045.72	1020	4.84
GA_OPP	75	11323.48	2760	**-2.17**	6688.93	1740	**-0.47**
MA(TS)	75	11714.65	1800	1.21	6923.67	1200	3.02
MA(SA)	75	11669.31	2040	0.82	6890.61	1260	2.53
MA(SHC)	75	11690.16	1560	1	6944.76	840	3.33
IH(TS)	75	14542.56	4740	25.64	8815.91	900	31.17
IH(SA)	75	14545.82	2520	25.67	8825.00	1440	31.31
IH(SHC)	75	14549.54	4740	25.70	8832.98	1740	31.43

Table 5.4: Computational results for PMSP - Scenario 1 (* - No optimal solution could be found within 3h. VOpt, optimal value of PMSP or the best value found within 3h; ObjV, solution value of the heuristic; TCPU, CPU time; RdOH, relative difference between VOpt and ObjV)

the averages of the least costs, the computation times for 10 problem instances and the relative differences between the exact solution and the solution found by the heuristics.

GA, GA_OPP and all three memetic algorithms improve on average the best solution found by the CPlex solver for both scenarios and possession cost equal to 25. However if the possession cost is equal to 75, then only GA_OPP can improve the solutions found by the CPlex solver. The CPU time of the heuristics vary from 3 to 50 minutes, but anyway far under the 3 hours that was needed on average by the CPlex solver. From Table 5.4 and Table 5.5 we conclude again that GA_OPP gives on average the best solution, followed by the MA(SA) and MA(TS).

Comparing the results of the memetic algorithms with the three local search algorithms, we observe that on average the MA(SA) has the highest computation time, followed by MA(TS), while the MA(SHC) uses the least time. Measured over both scenarios,

Method	PossC	SCENARIO 2					
		n = 25			n = 15		
		VOpt/ObjV	TCPU(s)	RdOH(%)	VOpt/ObjV	TCPU(s)	RdOH(%)
CPlex	25	10086.84*	>10800	-	5705.28	8462.13	-
OBH	25	10132.89	1.3	0.46	5816.55	1.3	1.95
MCWF	25	10283.31	1.6	1.95	6030.27	1	5.7
GA	25	10037.4	1440	-0.49	5702.92	900	-0.04
GA_OPP	25	9932.88	3000	**-1.53**	5680.4	1980	-0.44
MA(TS)	25	9940.42	1320	-1.45	5662.87	900	**-0.74**
MA(SA)	25	9938.75	1560	-1.47	5663.5	900	-0.73
MA(SHC)	25	9963.29	1140	-1.22	5688.35	600	-0.3
IH(TS)	25	11041.51	1380	9.46	6322.09	900	10.81
IH(SA)	25	11045.73	2640	9.51	6325.09	1500	10.86
IH(SHC)	25	11042.56	4740	9.47	6331.05	1920	10.97
CPlex	75	11710.15*	>10800	-	6889.13	7163.2	-
OBH	75	12505.64	1.4	6.79	7359.05	1.4	6.82
MCWF	75	12363.12	1.2	5.58	7811.35	1.5	13.39
GA	75	12052.14	1440	2.92	7104.31	900	3.12
GA_OPP	75	11607.02	3060	**-0.88**	6922.39	1980	0.48
MA(TS)	75	11841.97	1320	1.13	6980.62	900	1.33
MA(SA)	75	11811.56	1620	0.87	6959.36	900	1.02
MA(SHC)	75	11868.36	1200	1.35	7006.41	600	1.7
IH(TS)	75	14551.14	1380	24.26	8781.47	900	27.47
IH(SA)	75	14576.5	2640	24.48	8797.93	1500	27.71
IH(SHC)	75	14565.59	4740	24.38	8793.33	1920	27.64

Table 5.5: Computational results for PMSP - Scenario 2

the memetic algorithm with tabu search improves the results of the genetic algorithm with 1.38% on average on the cost of about 13% extra computational time. For the memetic algorithm with the simulated annealing these percentages are 1.6% and 24.6% respectively, while the memetic algorithm with steepest hill climbing improves the results of the genetic algorithm with about 1.2% whilst reducing computation time with 10%.

The iterative heuristics are for both scenarios among the slowest methods, especially if the number of works increases. The three iterative heuristics give very comparable results, implying that when performed on a random initial solution, the used local search optimiser does not make a large difference. The differences between the different memetic algorithms are slightly bigger than between the different iterative heuristics. Note that the iterative heuristic only faces randomly generated solutions, which can be compared with the initialization phase of the memetic algorithms. Therefore, the better performance of the memetic algorithms in comparison with the iterative heuristics are probably caused by the genetic part of the memetic algorithm. Anyway, the results with respect to the iterative heuristics lead us to conclude that these algorithms are not suitable for solving

PMSP. Burke and Smith (1999b), Burke and Smith (1999a), however report that the iterative heuristics (IH(TS), IH(SA) and IH(SHC)) provide reasonably good solutions for their maintenance scheduling problems, which is not anymore the case in Burke and Smith (2000). Nevertheless, all these articles clearly state that the memetic algorithm using tabu search as a local optimiser produces the best results in comparison with the GA, MA(SA) and MA(SHC).

Method	PossC	n = 25		n = 15	
		Loss	Gain	Loss	Gain
GA	25	282.1	2822.5	177.8	1460
GA_OPP	25	248.3	2925	222.7	1602.5
MA(TS)	25	251	2875	166.7	1482.5
MA(SA)	25	257.4	2872.5	192.8	1485
MA(SHC)	25	275.4	2867.5	188.6	1522.5
GA	75	407.6	8767.5	276.5	4672.5
GA_OPP	75	645.6	10117.5	377.2	5332.5
MA(TS)	75	384.3	8955	267.1	4717.5
MA(SA)	75	410.0	9000	293.9	4807.5
MA(SHC)	75	442.3	9097.5	280.6	4755

Table 5.6: Gains/losses achieved by clustering maintenance activities

In order to reduce the total possession cost and the inconvenience for the train operators and travelers, the maintenance works that do not exclude each other should be scheduled as much as possible together for the same time period. However, grouping mostly implies that one deviates from the originally planned execution moments and thus some maintenance actions are more often performed than originally planned. This involves costs as well. In Table 5.6 we present for different solution methods and averaged over 10 instances the gains and losses achieved by combining different maintenance activities with each other. Without loss of generality we consider here only the routine works and we assume that each routine work can be combined with all the other routine works. The losses are calculated as function of the maintenance costs and the number of time periods that the executions of a routine works are earlier planned than their cycle length requires. The gains are equal to the possession cost when all the routine works are performed separately minus the possession cost under the policy. Once the possession cost per time period rises from 25 to 75, the losses increase as well. The reason behind this is that the more expensive the possession per time period is, the more works are planned for the same time period. Thus, the executions of the works are planned much earlier than the cycle length requires. The losses incurred for too early executions of the maintenance work in order to combine works together is reduced from 14% of the reductions in the total possession cost to just 5%.

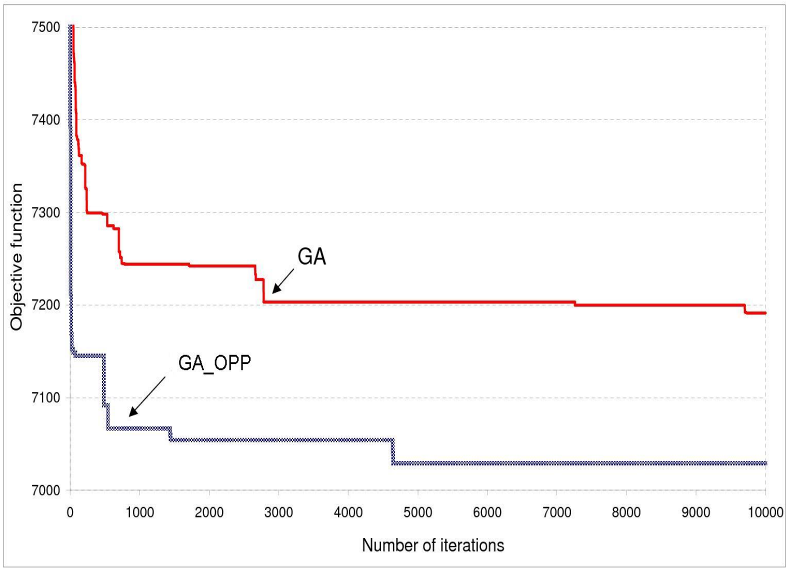

Figure 5.13: Progress of the GA and GA_OPP based on the best costs per iteration

In Figure 5.15 we present an example for a maintenance schedule made by the GA_OPP, that contains 15 routine works and 2 projects and with a planning horizon of 104 weeks (because of lack of space only 75 weeks are shown). All the routine works are allowed to be combined with other routine works or projects, but the projects cannot be combined with other projects. In the first 15 rows the routine works and in the last 2 rows the projects are scheduled. In order to minimize the possession costs the maintenance works are scheduled as much as possible together in the same time period. The Figure 5.15 clearly shows that the groups of maintenance works are not fixed during the planning period.

Furthermore, in Figure 5.13 and Figure 5.14 we compare the convergence of the GA with the GA_OPP and the convergence of MA(TS) with MA(SA) and MA(SHC), respectively, for the same problem instance that has been used for Figure 5.15. Comparing the convergence patterns of GA and GA_OPP in Figure 5.13, we observe that the initial solution is improved rapidly in the first hundreds of iterations. In the case of GA a major improvement in the solution value occurs approximately at iteration 3000 after which further improvements are marginal. In the case of GA_OPP the last major improvement occurs around iteration 4500. The solution does not change behind this point. In Figure 5.14 the initial solution is improved fast before iteration 200 and from that point very

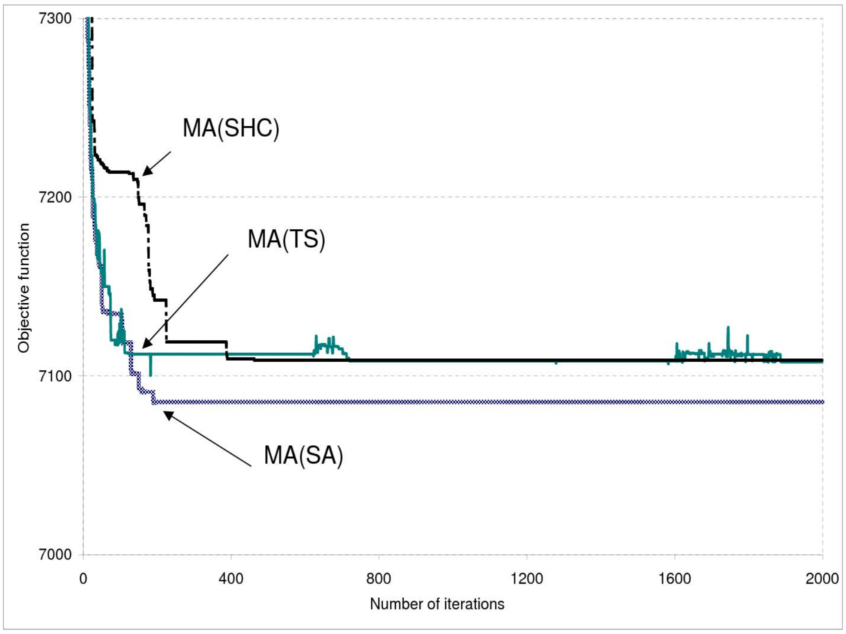

Figure 5.14: Progress of the memetic algorithm with the tabu search, simulated annealing and steepest hill climbing optimizers based on the best costs per iteration

little improvements in the solution value can be noticed. From the solutions found by the three memetic algorithms we can say the MA(SA) outperforms the MA(TS) which performers slightly better than SA(SHC). The best solution found by MA(SA) is however significantly improved by the GA_OPP.

In conclusion, we can say that the solution methods developed in this chapter, with the exception of the iterative heuristics, improved substantially the greedy heuristics from Budai *et al.* (2006). Moreover, on average GA_OPP improved the solution found by the CPlex solver for both scenarios and possession costs.

5.5 Conclusions

In this chapter our objective was to develop better techniques than the greedy heuristics (OBH and MCWF) developed in Budai *et al.* (2006) to solve the preventive maintenance scheduling problem (PMSP). We recall, that PMSP aims to assign (short) repet-

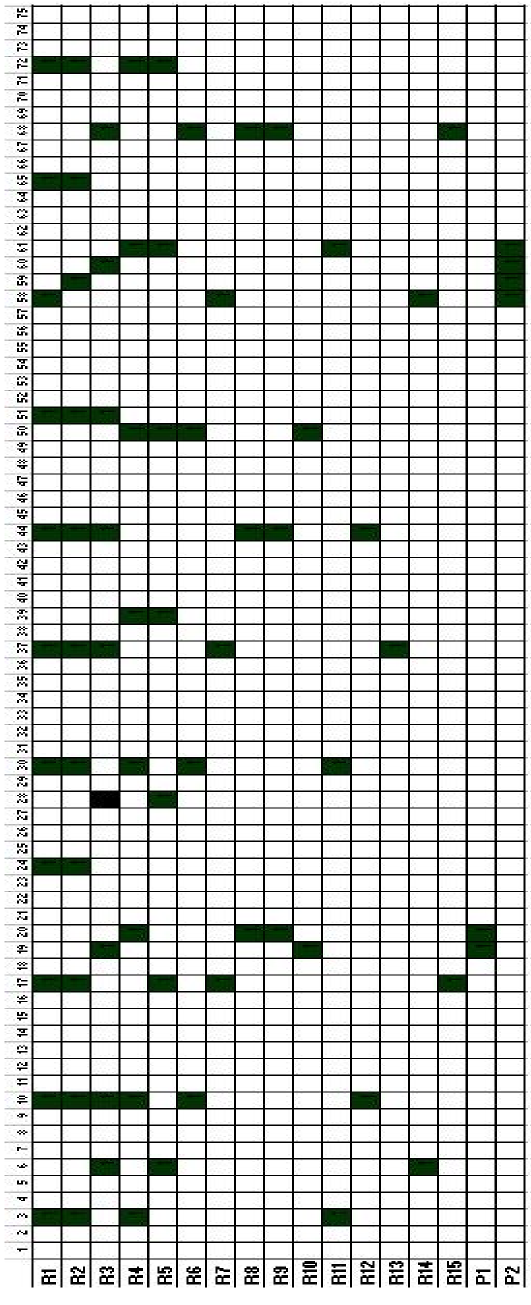

Figure 5.15: Example for a maintenance schedule containing 15 routine works and 2 projects

itive routine maintenance works and (long) unique projects to different time periods (months/weeks), minimizing the track possession cost and the maintenance cost.

In the literature several authors recommend using genetic algorithms (GAs) and memetic algorithms (MAs) for maintenance scheduling problems comparable to PMSP (e.g. Grimes (1995), Sriskandarajah *et al.* (1998), Burke and Smith (1997b), Burke and Smith (1999a), Burke and Smith (1999b), Burke and Smith (2000), *etc.*) Both algorithms are popular in maintenance applications because of their robust and fast search capabilities that help to reduce the computational complexity of large scale maintenance scheduling models. For this reason we implemented the genetic algorithm, memetic algorithm using three local search techniques (tabu search, simulated annealing and steepest hill climbing) and the iterative heuristic using the same local search techniques as the memetic algorithm for solving the PMSP. Moreover, we developed for the PMSP a two-phase opportunity–based heuristic (GA_OPP), where in the first phase, opportunities are created by using the GA and in the second phase all the executions of the preventive works are fitted as much as possible to these opportunities.

The genetic algorithm and the memetic algorithms with the three local search algorithms improve on average the solutions found by the CPlex solver for the instances where the possession cost is low (PossC = 25), while for high possession cost (PossC = 75) the solution found by the CPlex solver cannot be improved anymore, resulting in a relative difference between the solution value of the exact method and the solution value of the heuristic of 0.8% to 4.8%. Comparing the performance of the genetic algorithms with the performance of the memetic algorithms using the tabu search, simulated annealing and steepest hill climbing optimizers, we can conclude that the local search optimizers improve the genetic algorithm for almost all of the 80 instances. Within the memetic algorithm, the simulated annealing algorithm tends to give the best results in general, followed by the tabu search and finally by the steepest hill climbing. The iterative heuristic can be used to make comparisons, but it is not advisable to be used as a method to solve the PMSP, since it provides very poor results. The performance of the memetic algorithm is mainly caused by the genetic part, while the local search optimizers cause the improvement with respect to the genetic algorithm itself.

Furthermore, from the computational results we can conclude that GA_OPP can be used very well to solve PMSP, since on average GA_OPP improved the solution found by the CPlex solver for both scenarios and possession costs. Moreover, on average the GA_OPP outperformed the memetic algorithms with the three local search optimizers for 78% of the instances and the genetic algorithm for 94% of the instances, too. The success of the GA_OPP is due to the fact that it uses information about the structure of the problem. Given the opportunities the problem is much easier to solve. It also shows that standard GA's have problems in finding good solutions if only few paths in the solution space are possible. Furthermore, the solution value of the OBH and MCWF is on average significantly improved by the GA, GA_OPP and MAs, but not by the iterative heuristics. Comparing the computational times of these algorithms, we can conclude that OBH and MCWF are still the fastest heuristics, followed by the genetic and memetic algorithms. The GA_OPP needs somewhat more time than the GAs and MAs, but on average its CPU time (29-52 minutes) is still far below the computational time that the CPlex solver needed.

The ideas behind GA_OPP can also be extended to the maintenance planning on a network, especially the decomposition ideas. On a network level there is interest in whether adjacent single track grids are maintained at the same or adjacent time moments and whether there is a passage possible within the network. This can be taken into account when generating the opportunity structure by defining an appropriate cost structure. Fitting in the precise maintenance activities can be done on a second level in almost the same way.

Chapter 6

Rescheduling in Passenger Railways: the Rolling Stock Rebalancing Problem

Abstract

This chapter addresses the Rolling Stock Rebalancing Problem (RSRP) which arises within a passenger railway operator when the rolling stock has to be re-scheduled due to changing circumstances. The RSRP is relevant both in the short-term planning stage and in the real-time operations.

The RSRP has as input a timetable and a rolling stock schedule where the allocation of the rolling stock among the stations does not fit to the allocation before and after the planning period. The problem is then to correct these off-balances, leading to a modified rolling stock schedule that can be implemented in practice.

For practical usage of solution approaches for the RSRP it is important to solve the problem quickly. Since Maróti (2006) proves that RSRP is an NP-hard problem, this chapter focuses on heuristic approaches. We develop a two-phase heuristic approach for solving the RSRP and we compare it with the iterative heuristic approach developed in Maróti (2006). Both heuristics are described in this chapter and we compare them with each other on some (variants of) real-life instances of NS, the main Dutch passenger railway operator. Finally, to get some insight in the quality of the proposed heuristics, we also compare their outcomes with optimal solutions obtained by solving an existing rolling stock circulation model.

This chapter is based on Budai, Maróti, Dekker, Huisman and Kroon (2008).

6.1 Introduction

The rolling stock planning process of most railway operators is commonly divided into several planning stages. Huisman *et al.* (2005) distinguish four planning stages, namely strategic, tactical, operational and short-term planning. After these planning stages, the final plans are carried out and modified if necessary in the real-time operations. Strategic planning deals with long term decisions such as the acquisition of new rolling stock. At the tactical level, the different types of rolling stock are assigned to the different lines of the network. This is typically done once a year. The main goal of operational planning is to find a rolling stock schedule with low operational costs and high service quality. This schedule is basically to be carried out throughout the whole year. However, every day there are modifications to the timetable due to some extra trains or maintenance work on some parts of the railway infrastructure. These exceptions are handled during the short-term planning stage. The time horizon of short-term planning ranges from a couple of days to a couple of weeks.

This chapter deals with the *Rolling Stock Rebalancing Problem (RSRP)*, which is a problem faced in the short-term planning stage as well as during the real-time operations. The input consists of the timetable for a given planning period, the available rolling stock, the target inventories and an input rolling stock schedule, which is feasible except that it may contain some *off-balances*. An off-balance is defined as a deviation from the target inventory level of a certain type of rolling stock at a certain station. The primary goal is to construct a new rolling stock schedule with as few off-balances as possible. As secondary objective, other criteria related to costs, service, and robustness may be optimized as well.

To better understand the motivation for studying this problem, we first give some background information on rolling stock planning at NS, the main passenger railway operator in the Netherlands.

At NS, most trains are operated by electrical self-propelled *train units*, and only a few are operated by a locomotive and carriages. Therefore, we consider only train units in the remainder of this chapter. These train units are available in several types. Units of compatible types can be attached to each other to form longer compositions. Units of the same type are fully interchangeable.

The rolling stock schedule specifies the train *composition* of each train trip, *i.e.* how many units of each type are to be used for each timetable service and in which order. The practical feasibility of a schedule highly depends on the shunting possibilities of the stations. Since shunting is a complex problem on its own (see *e.g.* Freling *et al.* (2005), Lentink (2006)), NS uses an iterative approach. First, a rolling stock schedule is determined. Afterwards, when this schedule leads to infeasibility at certain stations, the schedule is modified. This process continues until there is an overall feasible solution. In practice, this may take several rounds. To speed up this process, several key aspects of the shunting process are taken into account during the creation of the rolling stock schedules. Examples are the restrictions on composition changes at certain stations: uncoupling (or coupling) of units can only take place at the appropriate side of the train.

During the operational and short-term planning stages, the process above is applied. The difference between both planning stages is that the time available to come up with a solution is much shorter in the short-term stage. The throughput time is even more binding for solving re-scheduling problems during the real-time operations, for instance due to major disruptions.

In recent years, NS introduced Operations Research based decision support tools for operational rolling stock planning (see Alfieri *et al.* (2006), Fioole *et al.* (2006), Maróti (2006) and Peeters and Kroon (2008)). However, the running time of these methods may reach several hours on instances that are smaller than typical re-scheduling instances. Thus the restrictive deadlines in re-scheduling raise the need for alternative approaches. Moreover, discussions with planners revealed that it is usually easy to come up with a rolling stock schedule that fulfills all requirements, except that the target inventories are not realized. This was our first motivation to study the RSRP.

A second motivation to study RSRP comes from disruption management (see Jespersen Groth *et al.* (2007)). During a disruption, the timetable is modified, and, as a consequence, the rolling stock schedule must be modified as well. As a result, the rolling stock units may not finish their daily duties at the location where they were planned to. This is not a problem if two units of the same type get switched. In many cases, however, the number of units ending up in the evening at a certain station differs from the number of units that has to start their next day's duty there. To prevent expensive dead-heading trips during the night, the rolling stock schedule must be modified such that the rolling stock is balanced before the night. This real-time version of RSRP is to a large extent equivalent to the RSRP in the short-term planning stage. The only difference is that the initial input schedule is not constructed by planners, but is caused by unexpected circumstances in the real-time operations. Moreover, in this case the off-balances only occur in the final inventories of the planning period.

Maróti (2006) showed that even the simple cases of RSRP are NP-hard. This fact, together with the need for fast solution approaches, motivated us to focus on heuristics. In this chapter we develop a two-phase heuristic approach for solving RSRP. Using (variants of) real-life problem instances of NS, this heuristic is compared with the iterative heuristic developed in Maróti (2006). Moreover, to get some insight in the quality of the heuristics, we also compare the results of the heuristics with the results of an existing optimal approach for operational rolling stock planning (see Fioole *et al.* (2006)). For the purpose of making the comparison of the two heuristics more clear, we also describe in this chapter the iterative heuristic.

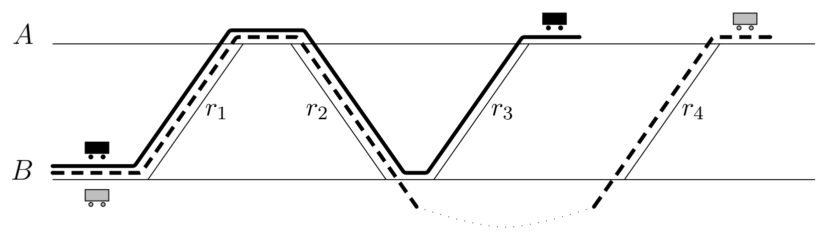

Figure 6.1: A time-space diagram indicating two stations, 4 trips, and two rolling stock units. The solid and dashed bold lines indicate the workload of the black and gray units, respectively.

The remainder of this chapter is organized as follows. In Section 6.2 we give a precise description of the RSRP. Section 6.3 contains a brief literature overview. In Section 6.4 we describe the two-phase heuristic for solving the RSRP. The iterative heuristic is described in Section 6.5. Computational results are discussed in Section 6.6. Finally, in Section 6.7 we draw some conclusions and we outline some directions for further research.

6.2 Problem description

In this section we define the RSRP in more detail. We are given the timetable for the planning period. The timetable defines a set of *trips*, each of which is characterized by a train number, departure and arrival times, departure and arrival locations as well as an estimated number of passengers. Moreover, we have a list of available rolling stock types and the number of available units per type. Figure 6.1 presents an example with stations A and B, with trips $r_1, \ldots, r_4$ and with a black and a gray rolling stock unit.

Next, we are given an *input rolling stock schedule* (or shortly *input schedule*), which assigns a rolling stock composition to each trip. The input rolling stock schedule satisfies the following requirements: the length of the composition on a trip r must be under a certain limit $\mu_r^{\max}$ (determined by the relevant platform lengths). Moreover, the trip has to be assigned at least a given number $\mu_r^{\min}$ of carriages in order to cover (a large part of) the passenger demand. A value of $\mu_r^{\min} = 0$ indicates that the trip may actually be canceled.

A rolling stock schedule can also be represented in terms of duties. A *duty* is the workload of a single rolling stock unit on a single day: it is a chain of tasks where a task is characterized by a trip and by the position of the unit in the composition of this trip, *e.g.* front, middle or rear.

The timetable also describes the rolling stock connections. Each trip (apart from arrivals in the late evening and from some rather exceptional trips) has a *successor*. In Figure 6.1, trip r_1 has as successor trip r_2, and trip r_2 has as successor trip r_3. A trip and its successor trip are in principle operated by the same rolling stock units. However, often one or more units can be coupled or uncoupled between the two trips. This is called a *composition change*. Since a trip is followed shortly – usually within minutes – by its successor trip, there is only a limited time for composition changes. A general rule states that coupling and uncoupling cannot be performed at the same time. Composition changes may take place at a pre-defined side of the train, either on the front side or on the rear side. This side is called the *shunting side* of the train at that station.

In Figure 6.1, the gray unit is uncoupled after trip r_2. Train units uncoupled from a train are not immediately available yet, only a certain *re-allocation time* ϱ later; this is to reserve time for necessary shunting operations before using the uncoupled unit in another train. A typical value for ϱ is 30 minutes.

We note that uncoupled units may be re-inserted later, like the gray unit being used for trip r_4. However, trip r_4 is not a successor of trip r_2, since the time between the two trips is too long. We can also see the effect of the shunting sides in Figure 6.1. Suppose the gray unit is running at the front position of the train on trip r_2, and suppose that station B allows shunting at the front of an arriving train. Then the black unit cannot be uncoupled after trip r_2.

A rolling stock schedule is *feasible* if (i) each trip is assigned a composition; (ii) the bounds on the train lengths are respected; (iii) composition changes (*i.e.* uncoupling and coupling of units) take place at the proper side of the trains; (iv) the uncoupled units are not used for any trip during the re-allocation time; and (v) the available amount of rolling stock is sufficient.

Next, we need the concept of the *inventory*. The inventory of a station at a given time moment is formed by the units that are currently staying idle at that station. These units can be coupled to a departing train, while uncoupled units go to the inventory. Note that in the inventory the order of the units is arbitrary. This is in contrast with the trains themselves where the order of the units is essential. We also note that, due to the concept of inventory, the compositions on the trips do not determine uniquely the duties: train units of the same type are indistinguishable in the inventory.

For each station, the *target initial inventory* is the number of units per type that are available there at the start of the planning period. The *target final inventory* is the number of units per type that are needed there after the planning period. For example, assume that the planning period is the entire Saturday and that there is no traffic at night. Then we can think of the target inventories for the rolling stock schedule of Saturday as the number of units that end at the stations on Friday evening, and that start at the stations on Sunday morning.

The input schedule may not comply with the target inventories. A station has a *deficit* (or a *surplus*) in the *initial* inventory of a given type if, according to the input schedule, the number of units of this type that are located at the beginning of the planning period at that station is higher (or lower) than the target initial inventory of this type. Similarly, a station has a *deficit* (or a *surplus*) in the *final* inventory of a given type if, according to the input schedule, the target final inventory of this type is higher (or lower) than the number of units of this type that are located at the end of the planning period at that station in the input schedule. The total *number of off-balances* in a rolling stock schedule is obtained by summing the surpluses over all stations and over all types. This number expresses how many units have to be involved in dead-heading trips, *i.e.* driving empty train units during the night in order to correct the off-balances in all stations. Since dead-heading trips are expensive, this number of remaining off-balances must be reduced as much as possible. Note that the total number of surpluses equals the total number of deficits.

The *Rolling Stock Rebalancing Problem (RSRP)* can now formally be defined as the problem of modifying the input rolling stock schedule during the planning period into a new rolling stock schedule such that (*i*) the new rolling stock schedule is feasible, and (*ii*) it contains a minimum number of remaining off-balances.

Note that if the number of off-balances is equal to zero, the new schedule satisfies the target initial and final inventory levels. Next to minimizing the number of off-balances, secondary objectives related to costs and service can be taken into account. In the experiments, we choose for an objective function which is a linear combination of the number of off-balances (with a very high weight), carriage-kilometers, shortage-kilometers and the number of composition changes. Carriage-kilometers express the operational costs of the railway operator. Seat shortage kilometers are computed by taking the expected number of passengers without a seat on a trip, multiplying it by the length of the trip and adding them up over all trips; the obtained value represents the service quality. The number of composition changes counts how many times units are coupled to or uncoupled from a train during a short stop. A schedule with a smaller number of composition changes is expected to be less sensitive to delays of trains.

The planning period may consist of several days. Then it is common to allow dead-heading trips every night during the planning period. The RSRP and the solution methods suggested in this chapter can easily be extended to such cases. In fact, we carried out computational tests on instances with a 2-day planning period.

Finally we note that operational rolling stock scheduling models described by Fioole *et al.* (2006) and by Peeters and Kroon (2008) can be easily extended for RSRP. However, their rather long computation times on large and complex problems motivated our research on heuristic solution approaches for solving RSRP.

6.3 Literature overview

A large number of publications addressed operational rolling stock planning, see Caprara *et al.* (2007) for a recent overview. We only mention here Peeters and Kroon (2008) and Fioole *et al.* (2006). Their models have basically the same specifications as those in this study. In the case when trains are not combined or split, Peeters and Kroon (2008) solve the problem by applying Dantzig-Wolfe decomposition and Branch-and-Price as solution technique. Fioole *et al.* (2006) extend the model for splitting and combining of trains. They use the commercial MIP software CPLEX to solve the model.

Compared to operational planning, literature on short-term railway rolling stock planning is scarce. Brucker *et al.* (2003) consider the problem of routing railway carriages through a railway network. The carriages should be used in timetable services or dead-heading trips such that each timetabled service can be operated with at least a given number of carriages, thereby satisfying the passenger demand. The order of the carriages is not considered. The objective is to minimize a non-linear cost function. The solution approach is based on local search techniques such as simulated annealing.

Ben-Khedher *et al.* (1998) study the short-term re-scheduling problem of the French TGV trains from a revenue management's point of view. The rolling stock circulation must be adjusted to the latest demand from the seat reservation system in order to maximize the expected profit.

Lingaya *et al.* (2002) deal with the effect of an altered timetable and passenger demand on the rolling stock schedules, focusing on the case of locomotive hauled carriages. They explicitly take the order of the carriages in the trains into account and assume that for each train a successor train has already been specified. Several real-life aspects, such as maintenance, are considered as well.

Substantial research has been carried out on aircraft and bus re-scheduling. Kohl *et al.* (2007) and Clausen *et al.* (2005) give overviews of airline disruption management, including a detailed list of aircraft re-scheduling publications and applications. The common solution approaches are based on multi-commodity network flows, thereby applying various exact and heuristic methods. Many of the models incorporate maintenance of the aircraft as well.

Recently, Li *et al.* (2007) introduced the single depot vehicle re-scheduling problem. It is motivated by the problem of updating bus schedules in the case when a single vehicle breaks down. The re-scheduling problem is formulated as a minimization problem over a number of vehicle scheduling problems.

A main distinguishing feature of railway (re-)scheduling is that trains may consist of multiple train units and that the order of the train units is to be regarded when they are attached to each other. In contrast, a single bus or aircraft is to be used for a flight or a bus trip. Also, strict airline maintenance regulations make it necessary to follow the path of each individual aircraft during the entire planning period. In railway re-scheduling, however, preventive maintenance is less binding, therefore train units of the same type can be considered as interchangeable.

We conclude that, although a large variety of related rolling stock scheduling problems has been described and partly successfully solved, railway rolling stock re-scheduling – in particular in the real-time operations – still lacks the appropriate models and solution methods.

6.4 A two-phase heuristic

In this section we describe a two-phase heuristic approach for solving the Rolling Stock Rebalancing Problem.

In Phase 1 we identify a number of elementary balancing possibilities (below abbreviated as BP) in the input schedule. These possibilities show how one unit (or several units) can be sent from a station with a surplus to another station with a deficit, such that the rolling stock balance remains unchanged for all other stations. The surplus and deficit are either in the initial or in the final inventory.

We restrict ourselves to three kinds of BPs. In some of the BPs, the excess unit of a certain type is coupled to a train and it will be uncoupled from the train once the station with a deficit in that type is reached. Thus, the train length on some trips is increased. In the second kind of BPs, the train length on some trips is decreased by removing one (or more) unit(s) of a certain type from a sequence of trips. In the third kind of BPs the duties of two different types of units are completely or partially exchanged. In Section 6.4.1 a more detailed description of some of the BPs is given.

A cost value is assigned to each BP expressing how the weighted sum of the carriage kilometers, seat shortage kilometers and shunting operations changes if the BP is indeed used. Dead-heading trips are also considered to be BPs. These correspond to unresolved off-balances, so their cost is defined as the penalty of an off-balance.

Given a set of all BPs that were defined in Phase 1, Phase 2 selects some of the BPs such that carrying out these selected BPs leads to a new rolling stock schedule without off-balances. Phase 2 minimizes the cost of the selected BPs and makes sure that they do not interfere with each other. This is done by solving an integer linear program.

A BP may solve off-balances only at the beginning or at the end of the planning period, but may also connect a surplus at the beginning of the planning period to a deficit at the end, or vice versa. In the latter cases, implementing that BP only would mean that one unit more or one unit less is used than in the input schedule. Constraints in the model of Phase 2 make sure that the output rolling stock schedule uses the same number of units as the input schedule does.

6.4.1 Phase 1: finding balancing possibilities

Here we show a couple of examples for balancing possibilities. Many other BPs can be computed using very similar ideas. These BPs can be obtained by examining the rolling stock duties and applying a number of straightforward constructive heuristics.

First we give examples for BPs where only one type of train unit is involved and after that examples for BPs with two different types of units are presented. The examples given for BPs with a single type of unit are divided into 4 categories, depending on whether off-balances are solved in the initial or final inventory. For the sake of simplicity, in each example presented below we assume that there is an off-balance of a single unit. However, these BPs are used in Section 6.6 also in the situations when stations have more than one off-balances.

6.4.1.1 Balancing possibilities with a single type of train unit

(i) Surplus and deficit in the initial inventory

Figure 6.2(a) shows a time-space diagram of a small railway system with two trips: r_1 and r_2 between stations A, B and C. The bold lines show the duties of the train units according to the input schedule. The target initial inventories are indicated by the gray train carriages. Both stations A and B have one unit available at the beginning of the planning period. Since the represented input schedule requires two train units to start at B, this schedule has an initial surplus of one unit at station A and an initial deficit of one unit at station B.

Figure 6.2: A balancing possibility for a case when station A has an initial surplus, and station B has an initial deficit: the train length on trip r_1 is decreased.

Figure 6.2(b) is a possible solution to the balancing problem. Trip r_1 from B to A is operated with one unit only; a second unit is attached to it at station A before leaving towards station C. Of course, this solution is a BP only if trip r_1 can be operated with a single unit, and if time and shunting capacity at station A are sufficient for carrying out the composition change (*i.e.* for coupling one unit) there.

Another example is shown in Figure 6.3, where the off-balance of stations A and B is resolved by increasing the train length on trip r_1 by one unit.

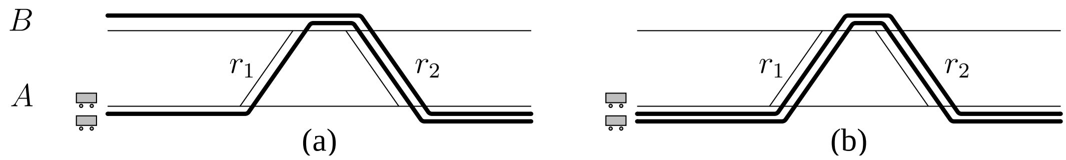

Figure 6.3: A balancing possibility for a case when station A has an initial surplus, and station B has an initial deficit: the train length on trip r_1 is increased.

(ii) Surplus and deficit in the final inventory

In Figure 6.4(a) there are two trips: r_1 and r_2 between stations A and B. The target final inventories are indicated by the gray train units. One train unit must be available at stations A and B at the end of the planning period. According to the input schedule both trips are operated with two train units and both units end their duties in station A. Thus, this schedule has a final surplus of one unit in station A and a final deficit of one train unit in station B. A possible solution to this balancing problem is shown in Figure 6.4(b). Trip r_2 from B to A is operated with one train unit only, the second train unit being uncoupled from the train at station B before leaving towards station A. This solution is a BP only if trip r_2 can be operated with a single train unit and if time and shunting capacity at station B are sufficient to uncouple one of the train units there.

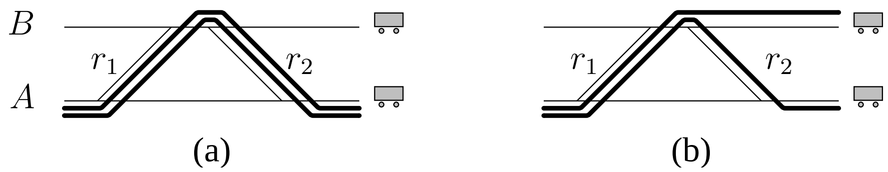

Figure 6.4: A balancing possibility for a case when station A has a final surplus, and station B has a final deficit: the train length on trip r_2 is decreased.

Other two examples of BPs are shown in Figure 6.5, where the final off-balance of stations A and B are resolved in two different ways. In the input schedule showed in Figure 6.5(a) two train units must be available at station B and one unit in station C at the end of the planning period. Since the represented input schedule has only one ending train unit in station B and one ending unit in station A, this schedule has a final surplus of one train unit in station A and a final deficit of one unit in station B.

Figure 6.5(b) and Figure 6.5(c) are two possible solutions to this balancing problem. In Figure 6.5(b) in order to resolve the surplus of A and the deficit of B, one has to modify the composition of two trips: r_2 and r_4 are operated by two units instead of one train unit each. Thus one train unit is uncoupled from the train at station C after trip r_2 and coupled later to the train leaving station C towards station B. The modified schedule has to agree with the shunting possibilities at station C and with the restriction that the uncoupled train unit after trip r_2 is available for trip r_4 only a certain reallocation time (usually 30 minutes) after being uncoupled.

In Figure 6.5(c) the off-balance is solved by modifying the composition of two trips: the train length on trip r_1 is reduced to a single train unit, while trip r_4 is operated with two train units, instead of one. Again, this is a solution to the balancing problem only if trip r_1 can be operated with a single unit and if time and shunting capacity at station C are sufficient for carrying out the composition change.

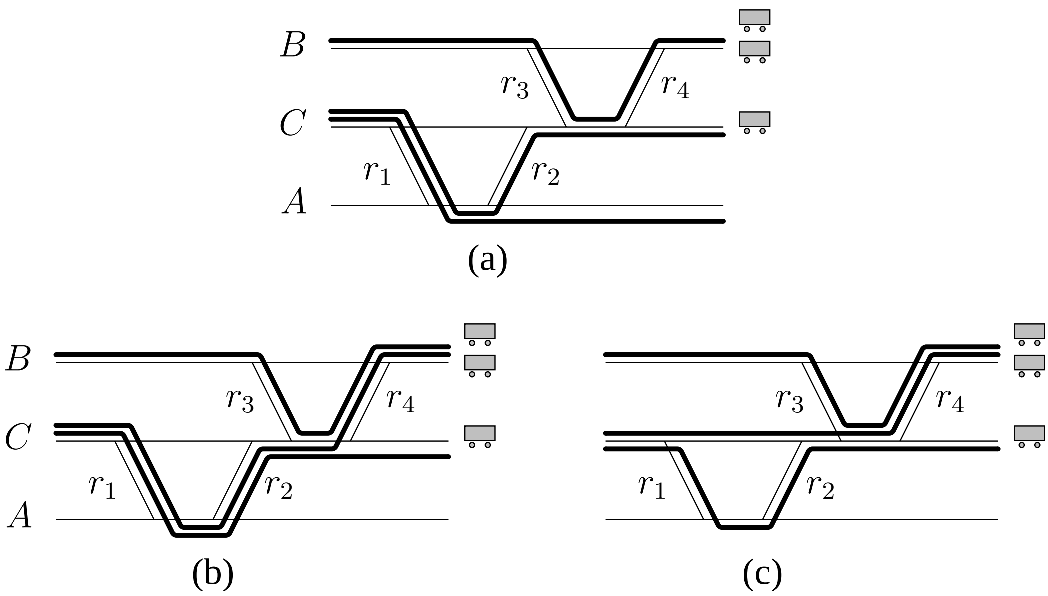

Figure 6.5: More elaborated balancing possibilities for a case when station A has a final surplus and station B has a final deficit.

(iii) Surplus in the initial inventory and deficit in the final inventory

In Figure 6.6(a) there are three trips: r_1, r_2, r_3 between stations A and B and each trip is operated with one train unit only. One unit each is available at stations A and B at the beginning of the planning period and at the end of the planning period one train unit must be available at both stations. Since in the input schedule no train units arrive at station B in the end of the planning horizon and the available unit in the initial inventory of station A is not used, this schedule has an initial surplus of one unit at station A and a final deficit of one unit at station B. A possible solution to this problem is presented in Figure 6.6(b), namely the length of the train on trip r_2 is increased by one unit and at station B one of the units is uncoupled from the train before it leaves towards station

A. This solution is a BP, if all the requirements with respect to shunting possibilities and train lengths are fulfilled.

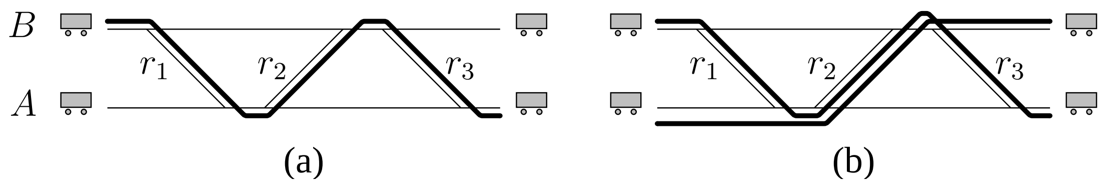

Figure 6.6: A balancing possibility for a case when station A has an initial surplus, and station B has a final deficit: the train length on trip r_2 is increased.

(iv) Surplus in the final inventory and deficit in the initial inventory

In Figure 6.7(a) trip r_1 is operated with two units and trip r_2 with a single unit. Moreover, one train unit is available at station B at the beginning of the planning period and one unit must be available at the same station at the end of the planning period. Since the represented input schedule requires two train units to start at B and at the end of the planning period no units must be available at A, this schedule has a final surplus of one unit at station A and an initial deficit of one unit at station B. A possible solution to this balancing problem is shown in Figure 6.7(b). In the modified schedule trip r_1 is operated with one unit only.

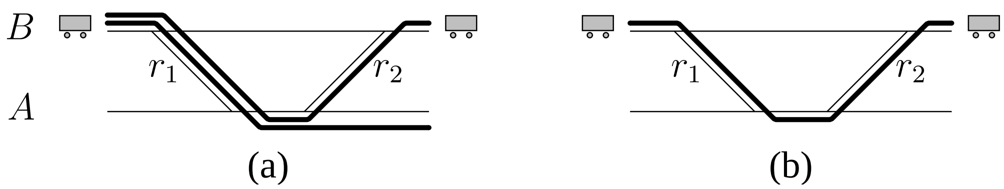

Figure 6.7: A balancing possibility for a case when station A has a final surplus, and station B has an initial deficit: the train length on trip r_1 is decreased.

6.4.1.2 Balancing possibilities with two different types of train units

More complicated BPs arise when solving at once two off-balances for two different types of train units. This can be done by switching two complete or partial duties of the input schedule. Such BPs are shown in Figure 6.8 and Figure 6.9.

Figure 6.8(a) shows a time-space diagram of a small railway system with six trips between stations A, B and C. These trips are operated with two types of train units: t and t'. The bold lines show the duties of the train unit of type t and the dashed lines the duties of the train unit of type t'. Moreover, the target initial inventories of type t

are indicated by the black train carriages and the target initial inventories of type t' by the gray train carriages. One unit of type t is available at station A and one unit of type t' at station B, both at the beginning of the planning period. Since the input schedule requires one train unit of type t at B and one train unit of type t' at A, this schedule has:

- an initial surplus of one unit of type t at station A and an initial deficit of one unit of type t at B and
- an initial surplus of one unit of type t' at station B and an initial deficit of one unit of type t' at A.

Figure 6.8(b) shows a possible solution to this balancing problem, namely switching in the input schedule the complete duties of the train units of type t and t'. Since both duties end in station C, switching the two duties does not have any effect on the final inventories. Of course, this solution is a BP only if trips r_1, r_2, r_3 can be operated with a train unit of type t and trips r_4, r_5, r_6 with a train unit of type t'.

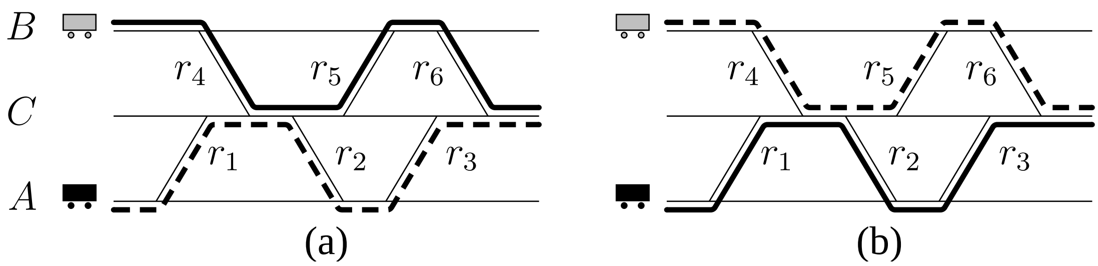

Figure 6.8: A balancing possibility for solving at once two initial off-balances for two different types of train units.

Another example is shown in Figure 6.9, where the off-balances are now in the final inventories. One train unit of type t' must be available at station A and one unit of type t at station B at the end of the planning period. In the input schedule a train unit of type t' ends in station B and the train unit of type t ends in station A. Therefore, this input schedule has:

- a final surplus of one unit of type t at station A and a final deficit of one unit of type t at B and
- a final surplus of one unit of type t' at station B and a final deficit of one unit of type t' at A.

In Figure 6.9(b) a possible solution to this balancing problem is given, namely after a common idle period at station C the train units of type t and t' take over each other's duty till the end of the planning period. Again, this solution is a BP only if the modified schedule agrees with the shunting possibilities at station C and with the restrictions with respect to re-using earlier uncoupled train units. Moreover, trip r_4 should be possible to be operated with a train unit of type t and trip r_2 with a train unit of type t'.

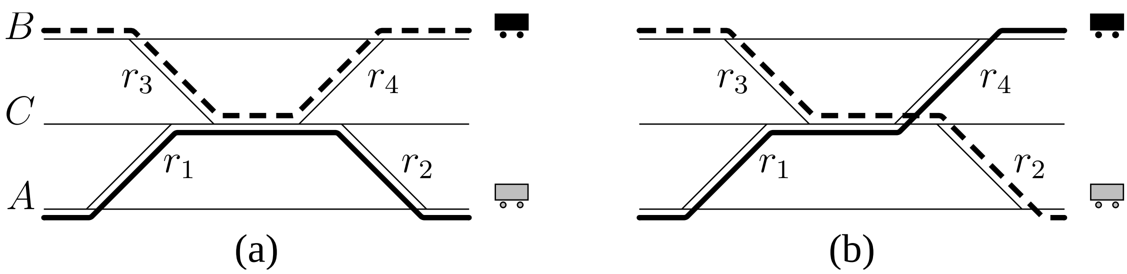

Figure 6.9: A balancing possibility for solving at once two final off-balances for two different types of train units.

6.4.2 Phase 2: combining the balancing possibilities

Given a set of all BPs that were defined in Phase 1, we choose in Phase 2 those BPs that minimize the sum of carriage kilometers, shortage kilometers, shunting movements (or composition changes) and number of dead-heading trips. The latter is equivalent with the number of off-balances.

The feasibility of a BP for a certain trip depends on the details of the composition of the trip, as well as on the compositions of its predecessor and successor trip. Therefore it is not possible to determine a priori whether certain combinations of BPs that modify the composition of the same trip result in a feasible rolling stock schedule. Thus, selecting several BPs may result in a conflict, e.g. by exceeding the maximal allowed train lengths. To stay on the safe side, we allow BPs to be selected simultaneously only if each trip gets modified at most once by the selected BPs. Thereby we guarantee that the selected BPs can be implemented in practice indeed. A disadvantage of doing so is that the solution space might be restricted too much, since BPs that modify the same trip might be used without conflicts in practice.

The BPs with overall minimum cost are selected with the following integer linear programming model.

Let E be the set of all BPs that were generated in Phase 1, S the set of all stations, T the set of train unit types and *Trip* the set of all trips. Let $b^{\text{beg}}_{s,t} \in \{0, \pm1, \pm2, \dots\}$ and $b^{\text{end}}_{s,t} \in \{0, \pm1, \pm2, \dots\}$ denote the surplus or deficit in the initial and final inventory of type $t \in T$ on station $s \in S$; a positive value indicates a deficit and a negative value indicates a surplus.

Let c_e be the cost of $e \in E$. Furthermore, we write $d^{\text{beg}}_{s,t,e}$ (or $d^{\text{end}}_{s,t,e}$) for how much the initial (or final) inventory of type $t \in T$ at station $s \in S$ is increased when applying an $e \in E$. Note that $d^{\text{beg}}_{s,t,e}$ and $d^{\text{end}}_{s,t,e}$ may be negative. The set Γ_e contains the trips that are modified by $e \in E$.

For each $e \in E$, let x_e be a binary decision variable expressing whether or not e is selected. Then the BP selection problem can be formulated as follows.

$$\text{minimize} \quad \sum_{e \in E} c_e x_e \tag{6.1}$$

$$\text{s.t.} \quad \sum_{e \in E} d^{\text{beg}}_{s,t,e} x_e = b^{\text{beg}}_{s,t} \qquad \forall s \in S, \forall t \in T \tag{6.2}$$

$$\sum_{e \in E} d^{\text{end}}_{s,t,e} x_e = b^{\text{end}}_{s,t} \qquad \forall s \in S, \forall t \in T \tag{6.3}$$

$$\sum_{e \in E:\, v \in \Gamma_e} x_e \leq 1, \qquad \forall v \in \mathit{Trip} \tag{6.4}$$

$$x_e \in \{0,1\} \qquad \forall e \in E \tag{6.5}$$

The objective function (6.1) minimizes the total cost of the selected BPs. At each station and for each train unit type, the sum of the changes in the initial (or final) inventory should be equal to the deficit or surplus in the initial (or final) inventory. This is ensured by constraints (6.2) (or (6.3)). Some BPs cannot be chosen together in the solution, since they use the same trip. This is expressed by constraints (6.4). Finally, constraints (6.5) state that the decision variables are binary.

In our computations, it turned out to be essential to strengthen the model by adding the following valid inequalities:

$$\sum_{e \in E: d^{\text{beg}}_{s,t,e} > 0} x_e \geq 1 \qquad \forall s \in S, \forall t \in T : b^{\text{beg}}_{s,t} > 0. \tag{6.6}$$

These constraints state that, if a station has a deficit in the initial inventory in a given type, then at least one BP is selected which increases the initial inventory of that type at that station. Furthermore, inequalities similar to (6.6) but $b^{\text{beg}}_{s,t} > 0$ replaced by $b^{\text{beg}}_{s,t} < 0$, $b^{\text{end}}_{s,t} > 0$ and $b^{\text{end}}_{s,t} < 0$ are also added to the model.

For each type t, the sum of $b^{\text{beg}}_{s,t}$ over the stations is zero. Therefore, if a model chooses some BPs connecting an initial surplus to a final deficit (i.e. where the BP increases the number of train units needed), then the appropriate number of BPs from an initial deficit to a final surplus will also be selected. So at the end, the updated rolling stock schedule uses the same number of train units as the input schedule.

Furthermore, we assume that E contains BPs corresponding to dead-heading (*i.e.* empty re-positioning of train units) between each pair of stations. Therefore the model (6.1) – (6.5) always has a feasible solution.

As we already mentioned before, the disadvantage of the two-phase heuristic is that it restricts the solution space, since the heuristic forbids two BPs to be selected both if there is a single trip that these BPs try to modify. Although it is very well possible that in practice both BPs fit together, it is hard to check this pairwise feasibility a priori. In order to compensate for this restriction, we apply the two-phase heuristic several times in a row until no further improvement is observed. That is, after each iteration, the BPs corresponding to dead-heading trips are deleted from the solution, and the two-phase heuristic is carried out once more. From that point of view, the method bears some similarity with column generation techniques: BPs are generated dynamically based on the current rolling stock schedule, and the "master problem" (6.1) – (6.5) selects the BPs into an overall solution.

The model (6.1) – (6.5) is solved by general purpose MIP software. The performance of the heuristic is presented in Section 6.6.

6.5 An iterative heuristic

In this section we describe an iterative heuristic approach for the Rolling Stock Rebalancing Problem. This heuristic has been already presented in Maróti (2006), but since the two-phase heuristic approach described in Section 6.4 is compared with this heuristic, we decided to describe the iterative heuristic in this chapter too.

In each iteration, either a *type switching step* or a *re-routing step* is carried out. Both steps intend to decrease the number of off-balances in a greedy way. The overall algorithm stops if no step can bring any further improvement.

Type switching steps consider pairs of rolling stock units of different types. The algorithm checks whether exchanging the duties of two units over the whole horizon results in a feasible rolling stock schedule and also whether the exchange decreases the number of off-balances. The two units whose exchange leads to the largest improvement are in fact switched, yielding an updated rolling stock schedule. Thereafter another iteration is launched. Type switching steps are straightforward, therefore they are not further described.

Re-routing steps attempt to solve a special case of the problem with *an off-balance of a single train unit*. We call this problem 1-RSRP. Maróti (2006) proved that it is an NP-complete problem to decide whether an instance of the Rolling Stock Rebalancing Problem has a feasible solution, even if only a single station has a surplus in the final inventory and another station has a deficit in the final inventory. Moreover, the problem remains NP-complete in the case of an off-balance of a single train unit.

Having found a solution for 1-RSRP, the rolling stock schedule is updated accordingly. The updated schedule has one off-balance less. Then another iteration is carried out. In the next section we describe the re-routing step in more detail.

6.5.1 The re-routing step

Here we assume that the input schedule realizes the target inventories except at two locations: there is a surplus of one unit of type t in the *initial* inventory of station A and there is a deficit of one one unit of the same type t in the *initial* inventory of station B. Note that deviations from the target *final* inventories of stations A and B can be handled in a very similar way. The goal is to modify the input rolling stock schedule in order to resolve these off-balances.

Discussions with planners revealed the desire to change the input schedule not too deeply when solving an off-balance of a single unit of a certain type t. This motivates the basic restriction in the heuristic approach to 1-RSRP: *the input schedule is to be modified in such a way that the circulation of every train unit type differing from t must not be changed.* So for example, if a trip has composition '*tab*' in the input schedule with train unit types a, b and t, then the algorithm must assign to this trip a unit of type a and a unit of type b in this order, and any number of units of type t before, between and after them. In particular, the modified composition can be '*ab*', '*atb*', '*attb*', '*tatbt*'*etc.* However, it cannot be '*taa*' or '*tba*' since those would change the circulation of types 'a' and 'b'.

The main idea of the re-routing step is to represent the problem as a single commodity network flow problem with side constraints. The side constraints express that (i) the train lengths lie between the given lower and upper bounds, and (ii) each shunting operation is either coupling of units or uncoupling of units. The underlying graph structure ensures that in case of composition changes, units are coupled to or uncoupled from the proper side of the trains.

The algorithm relaxes the side constraints, solves the network flow problem, and checks if the obtained flow satisfies the side constraints. If the side constraints are violated, the algorithm terminates. This approach is justified by our computational results where it turns out that none of the several thousand test runs leads to a network flow violating the side constraints. If, on other test problems, the side constraints were often violated, one would need to refine the network flow relaxation or to solve 1-RSRP as an Integer Program.

The graph representation

We represent the 1-RSRP as a network flow problem. To do so, we build up a graph $G = (V, E)$ which is a variant of a usual time-space network occurring in public transport problems. Let us start with an empty graph.

A time moment j is *relevant* at station C if a trip departs at j from C or if a trip r arrives at $j - \varrho$ at C where ϱ is the re-allocation time. In addition, the begin and the end of the planning period are also relevant. Create a *station node* for each pair (C, j) where C is a station and j is a relevant time moment at C. For each pair j, j' of consecutive relevant time moments at station C, draw a *station arc* from the node associated with (C, j) to the node associated with (C, j'). The flow values on the station arcs shall express the current inventories of type t at the stations. Station nodes at the beginning of the planning period are the *source nodes*, station nodes at the end are the *sink nodes*.

Consider a trip r and suppose that the input schedule assigns composition

$$\underbrace{t\ldots t}_{k_1^{(r)}}\, t_1\, \underbrace{t\ldots t}_{k_2^{(r)}} \ldots t_{\ell_r - 1}\, \underbrace{t\ldots t}_{k_{\ell_r}^{(r)}} \tag{6.7}$$

to trip r where $t_1, \ldots, t_{\ell_r - 1}$ denote train unit types differing from t. We assume that the left-hand side of this string corresponds to the front side of the train. That is, train units of type t are assigned to trip r in ℓ_r possibly empty groups, separated by $\ell_r - 1$ units of other types. The heuristic algorithm shall only modify the integer values $k_1^{(r)}, \ldots, k_{\ell_r}^{(r)}$, and leave the train units $t_1, \ldots, t_{\ell_r - 1}$ unchanged.

For example, if a and b represent train units of types different from t and a train consists of 4 train units in the composition $atbt$, then $\ell_r = 3$ and $k_1^r = 0$ and $k_2^r = k_3^r = 1$.

For each trip r, we create ℓ_r new nodes $u_1^{(r)}, \ldots, u_{\ell_r}^{(r)}$ corresponding to the groups of type t at the departure of r, and create ℓ_r new nodes $v_1^{(r)}, \ldots, v_{\ell_r}^{(r)}$ corresponding to the arrival of trip r. Moreover, we draw the arcs $u_i^{(r)} v_i^{(r)}$ for each $i = 1, \ldots, \ell_r$. We call these arcs *trip arcs*.

Let r' be the successor trip of trip r and suppose that in the input schedule, units are uncoupled from the arriving trip r. We also assume that the uncoupling takes place at the rear side of the train. Then our graph representation does not need to contain the possibility of coupling any unit to trip r' and we have $\ell_r \geq \ell_{r'}$. Physically, the train is split into two parts at a point that lies in the $\ell_{r'}{}^{\text{th}}$ group of the arriving composition. Then the first (*i.e.* left-most in (6.7)) $\ell_{r'} - 1$ groups go over unchanged to become the first $\ell_{r'} - 1$ groups of trip r'. The last (*i.e.* right-most in (6.7)) $\ell_r - \ell_{r'}$ groups (if any) are uncoupled. Units in the $\ell_{r'}{}^{\text{th}}$ group of trip r can go over to the $\ell_{r'}{}^{\text{th}}$ group of trip r' or they can be uncoupled. These possibilities are expressed by the arcs shown in Figure 6.10(a) for the case $\ell_r = 3$ and $\ell_{r'} = 2$. Notice that the re-allocation time ϱ is respected. The construction can easily be adjusted if uncoupling takes place at the front side of the arriving trip.

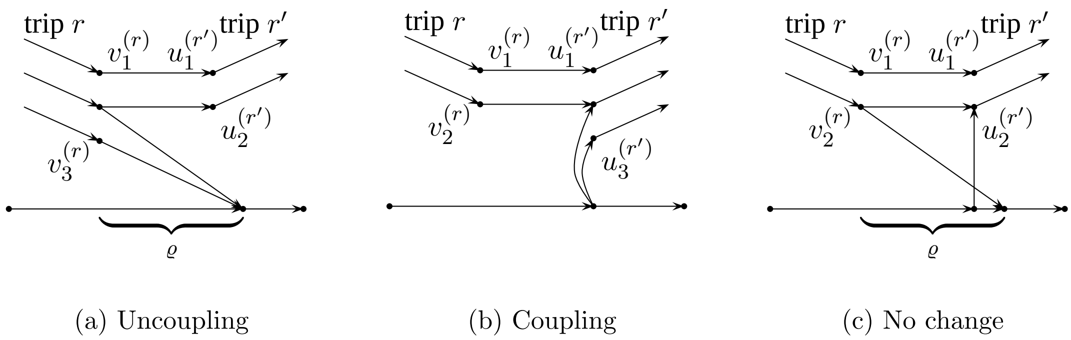

Figure 6.10: The graph representation of the cases when in the input schedule uncoupling, coupling or no composition change takes place between trips r and r'.

The cases when, according to the input schedule, units are added to the departing trip r' and when there is no composition change between trips r and r' are modeled similarly. Examples are shown in Figures 6.10(b) and 6.10(c).

We call an arc from a station node to a node $u_i^{(r)}$ a *coupling arc* and we call an arc from a node $v_i^{(r)}$ to a station node an *uncoupling arc* as they are intended to describe coupling and uncoupling of train units. Furthermore, a trip without a predecessor has to be supplied completely with rolling stock from the inventory at the involved station. Arcs that are similar to the coupling arcs described above are introduced for dealing with this situation. Similarly, for a trip without a successor, uncoupling arcs are introduced to allow the complete composition of the trip to be moved to the inventory at the involved station. This completes the definition of the graph G. Figure 6.11 indicates the graph representation for the black train unit type in a small railway network.

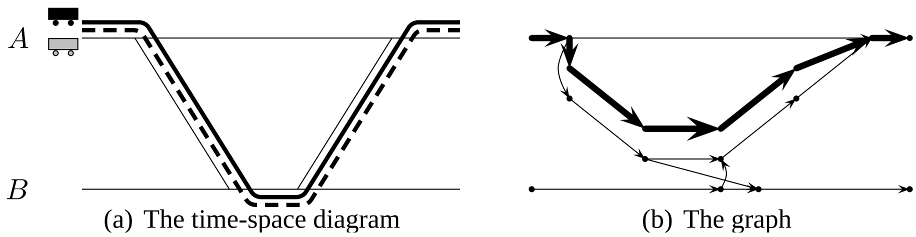

Figure 6.11: The graph representation of a small railway network and the flow of the black unit type: bold arcs have flow value one, other arcs have zero flow value.

Network flows in this graph

The movements of the units of type t in the input schedule correspond to a network flow x in $G = (V, E)$ as follows. Each trip arc corresponding to group i of trip r gets the corresponding value $k_i^{(r)}$. The number of coupled or uncoupled units of type t is assigned to the coupling and uncoupling arcs. The flow value on a station arc is the inventory of type t at that station during the time interval indicated by the arc. Then the source nodes have a (possibly zero) net out-flow, the sink nodes have a (possibly zero) net in-flow, and all other nodes satisfy the flow conservation law. Since this flow leads to off-balances in the initial or the final inventories of the train units, the flow is to modified so that the off-balances are solved.

The flow value on each arc is non-negative and, depending on the problem specification, it obeys certain upper bounds denoted by the capacity $g(a)$ of each arc a. For example, bounds on the station arcs may express the storage capacity of the stations. In addition, the following two side constraints (6.8) – (6.9) must be satisfied. First, the train length on each trip r obeys the lower and upper bounds $\mu_r^{\min}$ and $\mu_r^{\max}$:

$$\mu_r^{\min} \le \sum_{i=1}^{\ell_r} \lambda_t \times x\left(u_i^{(r)} v_i^{(r)}\right) + L_r \le \mu_r^{\max} \qquad \text{for each trip } r \tag{6.8}$$

Recall, that $\mu_r^{\min}$ and $\mu_r^{\max}$ represent the minimal and maximal number of carriages of the train on trip r. Furthermore, λ_t denotes the number of carriages of each train unit of type t, and L_r denotes the total number of carriages in those train units on trip r whose type differs from t.

Second, coupling and uncoupling may not take place at the same time between a trip r and its successor r':

$$\sum_{i=1}^{\ell_r} \sum_{\substack{a\in\delta^{\text{out}}(v_i^{(r)}):\\ a \text{ is an uncoupling arc}}} x(a) = 0 \quad \text{or} \quad \sum_{i=1}^{\ell_r} \sum_{\substack{a\in\delta^{\text{in}}(u_i^{(r')}):\\ a \text{ is a coupling arc}}} x(a) = 0 \tag{6.9}$$

where $\delta^{\text{in}}(v)$ (and $\delta^{\text{out}}(v)$) denotes the set of arcs entering (and leaving) node v.

Conversely, if a network flow in G satisfies side constraints (6.8) – (6.9), then it corresponds to a feasible rolling stock schedule.

In this section we assume that station A has an initial surplus of one unit and station B has an initial deficit of one unit. That is, the target final inventories are equal to the net in-flow of the sink nodes; the target initial inventories are equal to the net out-flow of the source nodes except for A and B. In order to resolve this off-balance, we have to find a network flow x' such that

$$\sum_{e\in\delta^{\text{out}}(A)} x'(e) = \sum_{e\in\delta^{\text{out}}(A)} x(e) + 1 \quad \text{and} \quad \sum_{e\in\delta^{\text{out}}(B)} x'(e) = \sum_{e\in\delta^{\text{out}}(B)} x(e) - 1$$

where we identified stations A and B with their source nodes. At each other node, the net in- and out-flow of x and x' must be equal. Furthermore, x' must satisfy the side constraints (6.8) – (6.9).

It is well-known in network flow theory that, if such a flow x' exists (without requiring (6.8) – (6.9)), then it can be obtained by modifying the flow x along an augmenting path P which is a directed $A - B$ path in the auxiliary graph G_x. The auxiliary graph G_x on the node set V is constructed as follows. Let G_x have a *forward arc* uv if $uv \in E$ with $x(uv) < g(uv)$. Let G_x have a *backward arc* vu if $uv \in E$ with $x(uv) > 0$. If there exists a directed $A - B$ path P in the auxiliary graph G_x, then the modified flow x' is defined as follows:

$$x'(uv) = \begin{cases} x(uv) + 1 & \text{if the forward arc } uv \text{ is used by path } P, \\ x(uv) - 1 & \text{if the backward arc } vu \text{ is used by path } P, \\ x(uv) & \text{otherwise.} \end{cases} \tag{6.10}$$

An arbitrary augmenting path P may lead to a violation of the side constraints (6.8) – (6.9). Actually, the feasibility version of 1-RSRP is NP-complete, therefore an augmenting path satisfying (6.8) – (6.9) cannot be found in polynomial time (unless P=NP). In our heuristic approach, we simply relax the side constraints (6.8) – (6.9): we look for an augmenting path and verify afterwards whether the updated network flow x' satisfies the side constraints (6.8) – (6.9).

If there is no augmenting path at all, then the instance of 1-RSRP is certainly infeasible. If there is an augmenting path and x' fulfills constraints (6.8) – (6.9) then the off-balance of stations A and B has been resolved. However, if there exists an augmenting path, but the side constraints are violated, then the algorithm reports that the off-balance could not be resolved. In the latter case, the answer might be wrong: other augmenting paths might result in satisfied side constraints. We note again that in our extensive computational tests, we did not find any augmenting path that lead to violated side constraints (6.8) – (6.9).

The algorithm as described above attempts to find any augmenting path. This reflects that the main objective is to resolve as many off-balances as possible. However, the additional secondary objective criteria (carriage-kilometers, seat shortage kilometers and the number of composition changes) can be taken into account by assigning cost values to the arcs of G. Then, according to classical network flow theory, arc cost in G_x are defined by $c_x(uv) = c(uv)$ if uv is a forward arc and by $c_x(vu) = -c(uv)$ if vu is a backward arc. Now we have to look for a minimum cost augmenting path in G_x.

6.6 Computational tests

In this section we report our computational results. All test instances are based on timetables of the so-called 3000 line of NS connecting Den Helder (Hdr) to Nijmegen (Nm). The stations are indicated in Figure 6.12. The line is operated in a cyclic timetable with a frequency of twice per hour in both directions. The timetable contains about 500 trips on each day. We studied instances for Saturday and Sunday, since these are the typical days of the week on which maintenance of the railway infrastructure takes place. This requires modification of the timetable and thus of the rolling stock schedules as well.

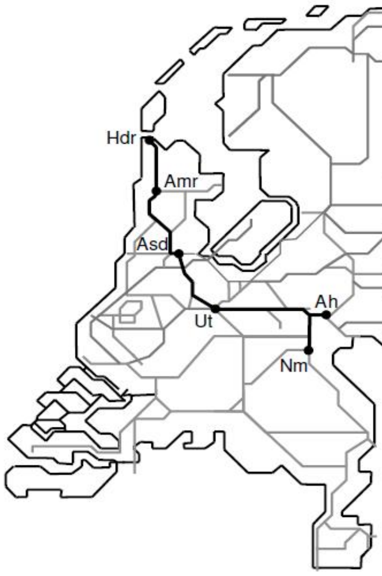

Figure 6.12: The 3000 line connecting Den Helder (Hdr) to Nijmegen (Nm) via Alkmaar (Amr), Amsterdam (Asd), Utrecht (Ut) and Arnhem (Ah).

Composition changes are possible at the terminals as well as at the intermediate stations Alkmaar (Amr) and Arnhem (Ah). Furthermore, units may start and finish their daily duties in Amsterdam (Asd) and Utrecht (Ut). The 3000 line is serviced by 11 units of type VIRM4 and 24 units of type VIRM6; these are double-deck units with 4 or 6 carriages, respectively (see Figure 6.13). The maximally allowed train length is 12 carriages, thus the VIRM types admit 7 possible compositions, namely 4, 6, 44, 46, 64, 66, and 444.

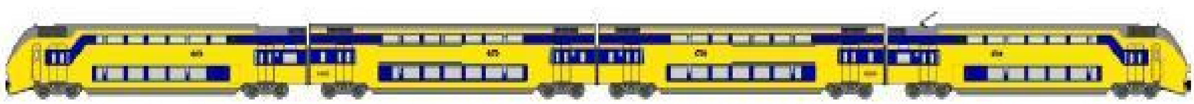

Figure 6.13: VIRM4 train unit

6.6.1 Instances

The timetable of the 3000 line contains about 500 trips on each day. We studied instances for Saturday and Sunday, since these are the typical days of the week on which maintenance of the railway infrastructure takes place. This requires modification of the timetable, and thus of the rolling stock schedule as well.

In the first computational tests we considered the timetable on Sunday and assumed that a certain part of the trajectory (either Amr-Asd or Asd-Ut or Ah-Nm) is closed either until 14:00 or for the entire Sunday. The reduced timetables have about 400 trips; in each of these six timetables, we updated the predecessor-successor pairs as well. The solutions are permitted to have off-balances on Sunday morning and evening.

Moreover, in the case that the infrastructure is blocked for an entire Sunday, it is common in practice to modify the rolling stock schedules both on the day itself and on the previous day. This leads to instances with a planning period of two days: Saturday and Sunday. The 2-day test problems concern about 900 trips.

To illustrate the behavior of the solution methods under different priorities, we considered three different settings for the relative importance of off-balances, carriage-kilometers, shortage-kilometers, number of composition changes. We refer to these settings as Obj-A, Obj-B and Obj-C. Table 6.1 contains the values of the coefficients in the objective functions. Besides heavily penalizing the remaining off-balances, Obj-A focuses on carriage-kilometers, Obj-B on composition changes, and Obj-C on seat shortages.

Table 6.1: Objective coefficients in the three cost structures.

Criterion	Obj-A	Obj-B	Obj-C
Off-balance	1,000.0	1,000.0	1,000.0
Carriage-kilometers	0.050	0.005	0.005
Composition changes	0.010	20.000	0.010
Shortage-kilometers	0.015	0.010	0.015

The VIRM4 and VIRM6 units have a limited number of possibilities to be attached to one another. Therefore in further artificial experiments, we changed the rolling stock types used. We split each VIRM4 and VIRM6 unit into two identical parts (*i.e.* VIRM2 and VIRM3). This results in as much as 48 possible compositions for each trip, increasing the complexity of the problem significantly. For these artificial rolling stock types, we considered the same one- and two-day instances and the same solution methods as for the original rolling stock types. All together, this results in 27 instances with VIRM4 and VIRM6 units, and 27 instances with VIRM2 and VIRM3 units. We refer to the instances with VIRM4/6 as V46 and to those with VIRM2/3 as V23.

6.6.2 Implementation issues

Throughout this section, Heur-1 denotes the two-phase approach described in Section 6.4 as well as its results, while Heur-2 denotes the iterative approach described in Section 6.5

as well as its results. To be more precise, the algorithm Heur-1 is composed of several consecutive runs of solving the the model (6.1) – (6.5), each run having the output of the previous run as input. The iterative process continues until no further improvement is observed. In all test cases this convergence occurred within 4 iterations.

For each of the instances, we computed the input schedule by the model of Fioole *et al.* (2006) with the objective functions Obj-A, Obj-B and Obj-C with the penalty of the off-balances being set to zero. Then we applied the heuristics Heur-1 and Heur-2 to resolve the off-balances. In order to be able to compare the solutions obtained by the heuristics with optimal solutions, we also solved the model of Fioole *et al.* (2006) with the penalties on the off-balances.

In the first phase of Heur-1 we identified about 20 classes of balancing possibilities and collected about 10,000–30,000 balancing possibilities. Thus the integer program in Phase 2 of Heur-1 has about 10,000–30,000 variables and 500–1,000 constraints. The graphs for Heur-2 have up to 5,800 nodes and up to 6,500 arcs.

The computations have been carried out on a PC equipped with a Pentium IV 3.0GHz processor and 1GB internal memory. For solving the model in Phase 2 of Heur-1 and for solving the model of Fioole *et al.* (2006) we used CPLEX 9.0 with the modeling software ILOG Opl Studio 3.7. The heuristic algorithms have been implemented in the C language (Heur-1) and in the Perl language (Heur-2).

6.6.3 The quality of the solutions

The results of the computational experiments of the heuristics Heur-1 and Heur-2 are presented in Tables 6.2 and 6.3. These tables also show the numbers of unresolved off-balances and the objective values in the optimal solutions obtained by CPLEX. These optimal solutions have the smallest possible numbers of off-balances.

Each row in the tables represents one instance. The instance names indicate the types used (V46 or V23) and the closed part of the infrastructure (C1 for Ah-Nm, C2 for Asd-Ut and C3 for Amr-Asd). Also, ‘H’ (or ‘F’) indicates that the planning horizon is Sunday and that the blockage takes the half of Sunday or (the full Sunday, respectively). The letter ‘W’ stands for the entirely blocked Sunday and for the planning horizon of the whole weekend. Finally, the objective function is added (A for Obj-A, *etc.*).

Tables 6.2 and 6.3 show that on many instances Heur-2 results in significantly less remaining off-balances than Heur-1. This can be explained partially by the fact that in Heur-1 two selected balancing possibilities may not touch the same trip, even if the technical and market requirements would allow using both of them. Yet, in some cases the greedy method in Heur-2 terminates with a higher number of off-balances than Heur-1. Moreover, Heur-1 appears to be able to balance the four optimization criteria better than Heur-2. Indeed, the contribution of carriage-kilometers, shortage-kilometers and composition changes to the objective function (this is the 'Rest' column in Table 6.2 and 6.3) is often higher for Heur-2 than for Heur-1. This is particularly true for the instances with Obj-C.

We can observe that the quality of the solutions highly depends on the structure of the input schedule. Table 6.4 summarizes how much more off-balance the heuristic methods produce compared to the optimal solutions; the average differences are referred to as Δ_1 and Δ_2. These values are much smaller for Obj-B than for Obj-A and Obj-C. In Obj-B, the number of composition changes has the largest weight. Therefore the input schedule for this objective contains a very small number of couplings and uncouplings. Then the heuristic methods find ways to resolve many off-balances. The input schedules for Obj-A and Obj-C are obtained by penalizing the carriage-kilometers and shortage-kilometers more heavily. The resulting larger number of composition changes is disadvantageous for both heuristic methods.

This can be explained partially as follows: if at a certain spot in the input schedule no composition change happens, then the models may allow both coupling or uncoupling of units there in the modified schedule. However, a coupling in the input schedule cannot be converted to uncoupling, and vice versa. Thus the more composition changes take place in the input schedule, the less flexibility is allowed for the models to solve the off-balances.

The input schedules of the V23 instances have on average 22.03 off-balances which is much more than the average of 11.15 of the V46 instances. However, the differences Δ_1 and Δ_2 of Table 6.4 are rather similar both for V23 and for V46. For Heur-1, the average differences in the number of off-balances are 4.41 and 5.30 for V46 and V23, respectively, while for Heur-2 these averages are 2.78 and 2.89.

That is, the heuristic algorithms perform relatively better on the V23 instances than on the V46 instances. This is not surprising: the shorter units give much more possibilities for adjustments without violating the constraints on the minimal and maximal lengths of the trains.

As said, Heur-1 is composed of several successive runs of the two-phase method of Section 6.4. In the first, second, and third run, the average differences Δ_1 over all V46 instances are 5.11, 4.44 and 4.41, while 6.74, 5.48 and 5.30 over all V23 instances. That is, the multiple consecutive runs indeed improved the performance of Heur-1. However, no further improvement was reached in a subsequent fourth run.

6.6.4 Computation times

Recall that the main motivation for using heuristic algorithms is the need for a quick solution process. Both heuristics are proved to be very successful in this respect, since algorithm Heur-1 has a running time of 4–5 seconds for each run and algorithm Heur-2 has a computation time of a 1–2 minutes for each run.

Table 6.2: Results for the V46 instances. For each instance, 'IOB' denotes the number of off-balances in the input schedule, 'OB' the remaining off-balance, 'Rest' the contribution of carriage-kilometers, seat shortages and composition changes, 'Obj' the objective value. 'Optimal' stands for the optimal solution obtained by CPLEX. For the optimal solution, 'ST' denotes the solution time (in seconds).

Instance		Heur-1			Heur-2			Optimal			
Name	IOB	OB	Rest	Obj	OB	Rest	Obj	OB	Rest	Obj	ST
46-C1-H-A	13	7	4,123	11,123	4	4,048	8,048	0	3,907	3,907	20
46-C1-H-B	8	2	449	2,449	2	624	2,624	0	545	545	24
46-C1-H-C	8	4	394	4,394	2	1,252	3,252	0	395	395	21
46-C1-F-A	11	7	3,771	10,771	8	3,707	11,707	3	3,660	6,660	15
46-C1-F-B	9	4	437	4,437	3	590	3,590	3	490	3,490	18
46-C1-F-C	13	5	996	5,996	5	3,772	8,772	3	379	3,379	15
46-C1-W-A	11	5	8,363	13,363	6	8,451	14,451	0	7,840	7,840	73
46-C1-W-B	9	4	1,184	5,184	1	1,875	2,875	0	1,165	1,165	74
46-C1-W-C	13	4	2,480	6,480	3	5,713	8,713	0	814	814	80
46-C2-H-A	11	4	4,048	8,048	3	3,747	6,747	0	3,654	3,654	19
46-C2-H-B	13	3	440	3,440	0	865	865	0	583	538	16
46-C2-H-C	12	4	381	4,381	3	1,766	4,766	0	371	371	19
46-C2-F-A	11	10	2,733	12,733	6	2,850	8,850	3	2,832	5,832	12
46-C2-F-B	11	5	328	5,328	4	516	4,516	3	428	3,428	12
46-C2-F-C	11	10	277	10,277	9	278	9,278	3	286	3,286	12
46-C2-W-A	11	8	7,546	15,546	3	7,917	10,917	0	7,088	7,088	103
46-C2-W-B	11	5	1,019	6,019	1	1,809	2,809	0	1,121	1,121	79
46-C2-W-C	11	8	2,375	10,375	6	3,268	9,268	0	728	728	86
46-C3-H-A	11	5	3,751	8,751	3	3,790	6,790	0	3,753	3,753	17
46-C3-H-B	12	2	456	2,456	0	849	849	0	557	557	17
46-C3-H-C	13	4	740	4,740	2	1,991	3,991	0	384	384	16
46-C3-F-A	11	7	2,931	9,931	6	2,935	8,935	4	2,961	6,961	12
46-C3-F-B	12	6	380	6,380	6	703	6,703	4	424	4,424	11
46-C3-F-C	11	7	298	7,298	7	774	7,774	4	302	4,302	11
46-C3-W-A	11	6	7,724	13,724	5	7,741	12,741	0	7,342	7,342	97
46-C3-W-B	12	7	1,066	8,066	2	2,142	4,142	0	1,178	1,178	71
46-C3-W-C	11	6	2,408	8,408	5	2,878	7,878	0	757	757	194

Table 6.3: Results for the V23 instances. Here we use the same notations as in Table 6.2.

Instance		Heur-1			Heur-2			Optimal			
Name	IOB	OB	Rest	Obj	OB	Rest	Obj	OB	Rest	Obj	ST
23-C1-H-A	23	6	4,007	10,007	2	3,928	5,928	0	2,961	2,961	730
23-C1-H-B	16	1	849	1,849	0	964	964	0	521	521	2,619
23-C1-H-C	23	10	1,276	11,276	5	6,568	11,586	0	302	302	310
23-C1-F-A	23	10	3,952	13,952	9	3,101	12,101	6	2,808	8,808	104
23-C1-F-B	22	7	440	7,440	6	762	6,762	6	451	6,451	133
23-C1-F-C	25	15	696	15,696	10	4,473	14,473	6	289	6,289	116
23-C1-W-A	23	4	7,561	11,561	3	7,894	10,894	0	5,824	5,824	715
23-C1-W-B	22	4	1,139	5,139	1	2,077	3,077	0	878	878	367
23-C1-W-C	25	8	2,350	10,350	4	5,805	9,805	0	595	595	1,173
23-C2-H-A	21	7	3,260	10,260	4	3,345	7,345	0	2,782	2,782	2,758
23-C2-H-B	20	2	522	2,522	2	870	2,870	0	542	542	247
23-C2-H-C	21	8	1,113	9,113	4	5,144	9,144	0	284	284	2,096
23-C2-F-A	24	10	2,851	12,851	10	2,171	12,171	6	2,128	8,128	114
23-C2-F-B	17	6	1,284	7,284	7	909	7,909	6	391	6,391	253
23-C2-F-C	28	16	218	16,218	10	408	10,408	6	217	6,217	114
23-C2-W-A	24	6	6,742	12,742	6	6,918	12,918	0	5,167	5,167	1,878
23-C2-W-B	17	4	1,469	5,469	2	2,299	4,299	0	782	782	333
23-C2-W-C	28	8	2,288	10,288	6	2,505	8,505	0	528	528	6,810
23-C3-H-A	20	3	3,634	6,634	5	3,224	8,224	0	2,917	2,917	348
23-C3-H-B	21	1	2,108	3,108	0	1,337	1,337	0	555	555	272
23-C3-H-C	24	7	314	7,314	4	2,265	6,265	0	301	301	392
23-C3-F-A	22	15	2,383	17,383	9	2,669	11,669	8	2,363	10,363	132
23-C3-F-B	18	9	340	9,340	8	990	8,990	8	831	8,371	111
23-C3-F-C	24	16	241	16,241	9	3,868	12,868	8	240	8,240	204
23-C3-W-A	22	7	6,855	13,855	5	7,628	12,628	0	5,404	5,404	597
23-C3-W-B	18	5	1,048	6,048	2	2,181	4,181	0	802	802	367
23-C3-W-C	24	7	2,309	9,309	5	8,186	13,186	0	551	551	1,264

Table 6.4: The average difference between the off-balances in the heuristic solutions (OB_1 for Heur-1 and OB_2 for Heur-2) and in the optimal solution (OB_{opt}).

		$\Delta_1 := OB_1 - OB_{opt}$	$\Delta_2 := OB_2 - OB_{opt}$
V46	Obj-A	5.44	3.78
V46	Obj-B	3.11	1.00
V46	Obj-C	4.67	3.56
V46	Total	4.41	2.78
V23	Obj-A	5.44	3.67
V23	Obj-B	2.11	0.89
V23	Obj-C	8.33	4.11
V23	Total	5.30	2.89

In Tables 6.2 and 6.3 we give the computation times of the exact optimization method of CPlex. Although this involves a relatively small instance of NS, it already shows how unpredictably the solution times grow when increasing the problem size. The V46 instances are easily solved for one-day planning period (these instances are denoted by 46-C*-H-* and 46-C*-F-* where * represents an arbitrary character) within 10–20 seconds. The two-day instances (denoted by 46-C*-W-*) require more than 5 times more CPU time on average. The V23 instances have a much more complex combinatorial structure due to the higher number of possible compositions. The solution time ranges from 2 minutes to 45 minutes, and a particular two-day instance requires nearly 2 hours.

6.7 Conclusions

In this chapter we formulated the Rolling Stock Rebalancing Problem (RSRP). This problem arises at various stages of the planning process of a passenger railway operator: from the short-term planning phase (*i.e.* planning some days or weeks ahead) till the real-time operations. Due to changes in the timetable (*e.g.* planned maintenance or unplanned disruptions) the previously created rolling stock schedules for a certain time period have to be adjusted.

A two-phase heuristic has been developed to solve the RSRP. The performance of this heuristic is compared with the performance of an earlier developed iterative heuristic for solving RSRP (see Maróti (2006)) and with the performance of the exact solution method of Fioole *et al.* (2006) used at NS, the main Dutch passenger railway operator. The comparison of the results is carried out on some (variants of) real-life instances of NS. These instances varied in size and complexity. For the purpose of making the comparison of the two heuristics more clear, we described in this chapter the iterative heuristic too.

From the results presented in Section 6.6 we can conclude that both heuristics are very fast, even if the problem size is increased by extending the planning period or by increasing the number of feasible compositions per trip. The results also show that both heuristics can be used effectively, not only for solving larger size planning problems, but they can also be used as a basis for solving real-time rescheduling problems in the case of a disruption of the railway system.

Chapter 7

Summary & Conclusions

This thesis presents a contribution to the field of operations research in two main areas, namely:

- *Maintenance scheduling*
- *Re-scheduling in the passenger railways*

In Part I of this thesis, which consists of Chapters 2, 3, 4, and 5, we deal with maintenance scheduling in line with production. First we look at the problem in general and after that we focus on applications in the railways. In Part II (Chapter 6) we deal with re-scheduling in the passenger railways and more precisely with the rolling stock rebalancing problem.

Below we give a summary on how this thesis fulfilled the stated aims presented in Chapter 1 in the two areas mentioned above. Moreover, we present for each chapter apart the main findings and conclusions.

Maintenance scheduling

In Chapter 2 we studied the literature on mathematical models that consider the relation between maintenance and production. This relation exists in several ways. First of all, maintenance is intended to allow production, but in order to perform the maintenance actions production is often stopped. Thus, in order to eliminate the negative effects of the production stoppage (e.g. high downtime costs, production losses), when planning maintenance one has to take the production needs into account. These needs are business sector specific and thus applications of planning models in different areas (railway, road, airline, electric power) have also been considered. Secondly, maintenance itself can also be seen as a production process which needs to be planned. Models for maintenance production planning mainly address allocation and manpower determination problems.

Finally, in production processes maintenance is mostly initiated by machine failures or low quality items. Since maintenance jobs take production capacity away, production and maintenance should therefore be planned in an integrated way. These integrated maintenance and production planning models determine optimal lot sizes while taking failure and quality aspects into account. Although many articles have been written on the interaction between maintenance and production, there is still a lack of studies on specific business sectors. As the demand for maintenance increases as public and companies are less likely to accept failures and the costs for maintenance are also likely to increase we do expect more research studies aimed at supporting maintenance and production decisions.

In Chapter 3 we continued to study the literature on maintenance planning on the railway infrastructure and especially on the effect of maintenance on the railway capacity. The resulting review is based on a general structuring of the maintenance planning process in eight phases. The following phases were distinguished: (1) Budget determination, (2) Long-term quality prediction, (3) Project identification & definition (diagnosis), (4) Project prioritization and selection, (5) Possession allocation and timetabling of track possession, (6) Project combination, (7) Short term maintenance and project scheduling, (8) Work evaluation and feedback loop. We could not find contributions to each planning phase, but the papers that we found are divided into two groups. The first group incorporates deterioration modelling (phase 2), defines the resulting work (phase 3) and makes some scheduling. The second group takes the maintenance work as input and tries to fit it in the timetable and makes a good manpower and equipment schedule. From this literature study we could conclude that scheduling of the preventive maintenance works on the railway infrastructure is very difficult, since there are many constraints to be considered. Firstly, carrying out maintenance on the rail infrastructure usually involves many disturbances for travelers (*e.g.* delays, canceled trains), and vice versa, the train operation restricts the length and the frequency of the infrastructure possession. Moreover, due to a couple of severe accidents in the last years the safety regulations for the track workers has become very very strict in the Netherlands. Furthermore, railway infrastructure maintenance costs have increased substantially in the past years too. All these facts show that there is a need for developing mathematical models and techniques that can assist the maintenance managers in making good (that is cost-effective) schedules for carrying out different maintenance activities on the railway infrastructure.

In Chapter 4 we introduced the Preventive Maintenance Scheduling Problem (PMSP), where a schedule for the repetitive routine works and once-only projects has to be found for one link such that the sum of the possession costs and the maintenance costs is minimized. The possession costs are mainly determined by the possession time, i.e. the time that a track is required for maintenance and cannot be used for railway traffic. In order to reduce track possession costs the maintenance activities are clustered as much as possible in the same period. We presented two versions of the PMSP, one with fixed intervals between two consecutive executions of the same routine work (this is called Restricted Preventive Maintenance Scheduling Problem (RPMSP)) and one with only a maximum interval (PMSP). The contribution of this chapter is twofold. First of all, we provided a mathematical programming formulation for both the PMSP and the RPMSP and we proved that PMSP is NP-hard. Secondly, since the maintenance scheduling problem is a complex optimization problem and for a large set of instances it was difficult and time consuming to solve the problem to optimality, it was necessary to develop some approximation methods, which still give solutions close to the optimal ones. In total four heuristics have been developed and the performance of these heuristics were compared with the optimal solution using some randomly generated instances. The tests showed that PMSP resulted in substantially lower overall costs than RPMSP, but the computation time was much longer and sometimes it was even impossible to get an optimal solution within 3 hours. Furthermore, the heuristics proved to be very fast (1–2 hundredths of seconds), however only two heuristics gave acceptable results in comparison with the CPlex solutions. Nevertheless, we could conclude that if the planners want to have a good (close to optimal) schedule in very short time, then either RPMSP or one of the heuristics can be used effectively.

In Chapter 5 we continued to study the Preventive Maintenance Scheduling Problem (PMSP). Our objective was to develop better techniques than the greedy heuristics developed in Chapter 4 to solve the PMSP. In the literature we saw that meta-heuristics were used with success for maintenance scheduling due to their robust and fast search capabilities that help to reduce the computational complexity of large scale maintenance scheduling problems. Based on this fact, in Chapter 5 we implemented a genetic algorithm, a memetic algorithm using three local search techniques (tabu search, simulated annealing and steepest hill climbing) and an iterative heuristic using the same local search techniques as the memetic algorithm for solving the PMSP. Furthermore, we also developed a two-phase opportunity–based heuristic, where in the first phase, opportunities are created by using the genetic algorithm and in the second phase all the executions of the preventive works are fitted as much as possible to these opportunities. From the computational results we concluded that excluding the iterative heuristics, that gave very poor results, the other solution methods improved on average the solutions found by the CPlex solver for the instances with low possession cost, while for the instances with high possession cost only the GA_OPP could improve the CPlex solution. Comparing the performance of the genetic algorithms with the performance of the memetic algorithms using the tabu search, simulated annealing and steepest hill climbing optimizers, we concluded that the local search optimizers improve the genetic algorithm for almost all of the instances. Within the memetic algorithm, the simulated annealing algorithm gave the best results in general, followed by the tabu search and finally by the steepest hill climbing.

Summarizing our findings for the maintenance scheduling problems we can say that the mathematical models and techniques developed in Part I of this thesis can assist maintenance managers in making in a reasonably short time cost-effective schedules for carrying out different preventive maintenance activities on the railway infrastructure. Moreover, the models presented in chapters 4 and 5 are just basic models, but they can be extended to solve many types of practical problems, since in reality there are many more constraints that a maintenance planner has to take into account. Finally, we would like to mention that the mathematical models and solution techniques developed for scheduling railway maintenance can also be applied for maintenance scheduling in other public/private sectors as well.

Re-scheduling in the passenger railways

Chapter 6 focused on rescheduling of the rolling stock in the passenger railways due to changing circumstances and more precisely on the Rolling Stock Rebalancing Problem (RSRP). Due to changes in the timetable (e.g. planned maintenance of the railway infrastructure or unplanned disruptions of the railway system), the input rolling stock schedule for a certain planning period contains a number of off-balances. RSRP is the problem of modifying the input rolling stock schedule such that the number of remaining off-balances is minimal, in combination with other criteria related to efficiency, service and robustness. Since RSRP arises in the short term as well as in the real-time planning phase, there is a need for fast solution methods. In Chapter 6 a two-phase heuristic method has been developed to solve RSRP. The performances of this heuristic is compared with the performance of an iterative heuristic developed in Maróti (2006) and with the performance of the exact solution method of Fioole *et al.* (2006) that is used at NS, the main Dutch passenger railway operator. The comparison of the results is carried out on some (variants of) real-life instances of NS. For the purpose of making the comparison of the two heuristics more clear, we also described in Chapter 6 the iterative heuristic. From the computational results we concluded that both heuristics are very fast and they can be used effectively, not only for solving larger size short-term planning problems, but also as a basis for solving real-time rescheduling problems in the case of a disruption of the railway system.

If we look at the three objectives of this thesis (see Section 1.4), we can conclude that all of them are fulfilled: (1) we reviewed the literature on maintenance planning in relation with production, on railway infrastructure maintenance planning and on rolling stock (re)scheduling, (2) for some of the identified problems we developed models and solution techniques and (3) the considered models and solution approaches were tested on real-life or randomly generated data.

Finally, we would like to provide some suggestions for future research on both areas discussed in this thesis.

1. In this thesis we showed that track possession planning is a very complex problem since several constraints (e.g disturbances for the travelers, length of the possession restricted by the train operation, *etc.*) should be taken into account. For this reason, we believe that the ideal planning process would be to schedule maintenance activities at the same time as constructing the train schedule. Instead of constructing and optimizing a train plan to minimize delays due to train conflicts and unplanned events, a train plan can be constructed to maximize time windows available for possible maintenance activities.

2. In order to produce a good maintenance plan we suggest to develop mathematical models that combine infrastructure deterioration with possession planning.

3. The models and techniques used for preventive maintenance scheduling problem can be easily extended by taking the maintenance resources (manpower, equipment, tools, spare parts) into account. The schedule is then developed based on available resources and job priorities.

4. We suggest to continue the research on the rolling stock rebalancing problem and develop solution methods that can deal with real-time rescheduling problems in the case of a disruption of the railway system.

Appendix A

Glossary of Terms

During this thesis we used many times some railway related terminologies that without a glossary is very difficult to understand, e.g. crossings, ballast cleaning, track possession time, etc. Furthermore, within a large project with partners from several countries, as "IMPROVERAIL" project (see Improverail) was, one of the most important tasks is to use a predefined common terminology, in order to assure a clear understanding.

Some glossaries have already appeared about railway systems. However, these contain solely the definitions of the railway components without any background information about railway maintenance. Therefore, the aim of this appendix is to provide a reliable terminology definition together with a description of the underlying maintenance procedures. We provide sometimes few pictures in order to be more clear.

ASSETS OF THE RAILWAY INFRASTRUCTURE

- Track: rails, ballast, sleepers, fastening;
- Switches;
- Crossings;
- Power supply/Catenary system;
- Signalling systems;
- Level crossings;
- Tunnels;
- Bridges.

BALLAST[1]

The purpose of ballast is to support the rail/sleeper combination, to distribute the load applied to it, to facilitate drainage to the track and thereby keep water away from rails and sleepers. The ballast is made up of stones of granite or a similar material and should be rough in shape to improve the ability of locking of stones (in this way they will resist better to the movements). The type of stones, which are used, can differ per country.

Ballast degrades as it becomes clogged with fine particles, which reduce its ability to fulfill the above mentioned functions. Such material accumulates within the ballast from several sources. In part, it is a product of natural deterioration, but there are significant usage-related effects. Abrasion occurs within the ballast itself and with the sleepers each time the track deflects under axle loads. Additionally, ballast also wears through friction each time it is subjected to track maintenance activities such as tamping. The passage of traffic may introduce other external pollutants such as brake dust and detritus from wagons, which might add to maintenance activity The functionality of ballast is maintained by ensuring that it remains relatively free from pollution by such fine materials. Maintenance and renewal works therefore include replacement and cleaning of track ballast. Ballast cleaning (see *Ballast cleaning*) involves removal, screening and return of the stones to the track and may be carried out manually or mechanically. Apart from regular tamping, (see *Ballast tamping*), the ballast has to be cleaned or replaced every few years. Other track forms include slab (non-ballasted) track, which does not require ballast (see *Slab track*).

BALLAST CLEANING[2]

The process for renewing the roadbed involves removing the ballast, cleaning it, and replacing it. The shaker screen separates the good ballast from the bad ballast. Clean ballast is then returned to the track for tamping along with more ballast to make a good roadbed. The small ballast is screened out, because it does not drain well. This process gives the tracks and sleepers a good foundation.

BALLAST TAMPING[3]

Ballast tamping is the process by which ballast is packed around the sleepers of a track to ensure the correct position for the location, speed and curvature. This can be

[1] Office of the Rail Regulator, http://www.rail-reg.gov.uk/upload/pdf/89.pdf, last accessed on 5 January 2009

[2] Esveld (2001)

[3] Railway Technical Web Pages, http://www.railway-technical.com/newglos.shtml, last accessed on 5 January 2009

done manually or mechanically by special tamping machines.

BRIDGE

A bridge carries a road or railway across a river, a valley or another obstacle. Therefore, it carries the direct loads imposed by rail traffic. The bridge deteriorates by ageing, corrosion and other 'environmental' factors. Deterioration and economic life of a bridge significantly depend on the intensity of use. Metal bridges are particularly liable to fatigue. This leads to frequent component replacement or even full-scale reconstruction of the bridge.

CATENARY SYSTEM

The catenary system consists of the overhead power cables for electric trains. While age and corrosion are important in the speed of degradation of catenary system, heavily used systems are more likely to require higher levels of maintenance and renewal.

CHAIR (or TIE PLATE)[3]

A chair is the cast steel component on a sleeper, which fixes the rail in the correct position. Depending on the design, the rail is secured to the chair by a clip (see *Clip*) or key (see *Key*).

CLIPS[4]

It is a "G" shaped piece of metal, used to attach the rail to the sleeper with a specific type of chair (see *Chair*).

CORRIDOR

A track corridor is a rail path connecting two major points within a railway system.

CROSSING (or FROG)[2]

A crossing is part of the track that allows two tracks to intersect at the same level.

CROSSOVER[4]

Crossover is formed from two switches that allow a train to switch from one track to a parallel track. The types of crossovers are presented in Figure A.1.

FASTENING[2]

[4]Glossary of Signalling Terms, http://broadway.pennsyrr.com/Rail/Signal/glossary.html, last accessed on 5 January 2009

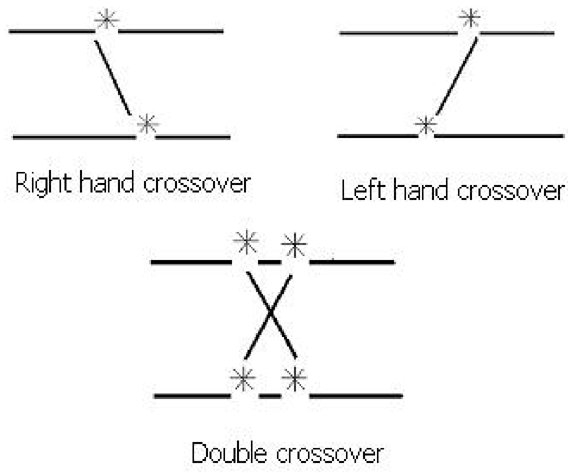

Figure A.1: Crossover types

The "fastening system", or shortly "fastenings" includes all the components, which together fix the rail to the sleeper. The general functions of the fastenings are to absorb the rail forces and transfer them to the sleeper, to damp vibrations caused by traffic as much as possible, to fix the rail in correct position.

FROG See *Crossing*.

GAUGE[4]

The distance between the inner faces of two rails on the same track is called gauge. In most of the railways, the gauge is 143.5 cm, although different countries may have its own "standard gauge".

HIGH SPEED LINE-SOUTH[5]

The High Speed Line-South (HSL-South) is a high-speed train service that links Amsterdam to Brussels, London and Paris.

KEY[6]

The key is a wedge of hard wood or steel inserted between rail and chair to hold rail firmly in position at correct gauge.

[5]http://www.hslzuid.nl/hsl/uk/hslzuid, last accessed on 5 January 2009

[6]Glossary of Model Railroad and Prototype Railroad Terms, http://www.hrtrains.com/glossery.html, last accessed on 5 January 2009

LEVEL CROSSING[6]

A level crossing is a crossing of a railway and a road at the same level.

LINE

A line consists of one or more adjacent running tracks forming a route between two points.

LINE CAPACITY[3]

Line capacity is the maximum possible number of trains capable of being operated over a line in one direction. It is usually expressed in trains per hour.

POINTS

See *Switches*.

POSSESSION[3]

When a section of track is required for maintenance and it is therefore blocked from train traffic, it is handed over by the operators to the engineers, who take "possession" of the track. When the track is returned to the operators, the engineers "give up possession".

RAIL[1]

Rail is a rolled steel shape designed to be laid end-to-end in two parallel lines on sleepers to form a track for railroad rolling stock. Rails deteriorate through natural processes, such as corrosion. In certain environmental conditions (such as heavy pollution, wet tunnels and salt-laden atmospheres) this process can be accelerated in such of way that it becomes the primary factor of maintenance and renewal. The life of the rail is also determined by fatigue. It is considered to be partly "consumed" each time it is subjected to the wheel loads of passing trains. As the cumulative traffic carried by a piece of rail increases with its age in service, there will be an increasing likelihood that the rail metal will develop internal defects. Monitoring of rail defects, and the actions necessary to respond to them when they are detected, are both important elements of rail maintenance.

RAIL GRINDING[2]

Rail grinding is the process to maintain a predetermined profile on the head of the rail in order to maximise rail life and minimise rolling resistance reducing wheel wear and improving fuel economy.

RAILROAD

A railroad represents a specific length of track.

RAILWAY

A railway is a transport system in which trains run on steel rails. On most railways, the tracks consist of two rails, which are placed exactly 143.5 cm apart.

RAILWAY NETWORK

Railway network means all railways in a given area.

ROLLING STOCK

Rolling stock consists of locomotives, passenger and freight vehicles owned or operated by a company.

SIDING[6]

An auxiliary track to a main or secondary track for the meeting or passing of trains is called a siding.

SIGNALS[3]

Signals are the visual indications passed to a train driver to advise the speed, direction or route of the train. There are almost as many types of signals as railways, but they fall into the following main categories:

- Hand signals - used mainly where there are no fixed signals or where the fixed signalling has failed. Generally, each railway has its own defined handsignals recognized by its operators.
- Semaphore signals - a fixed lineside signal where the stop indication is displayed as a horizontally positioned arm and proceed as a 45 or vertical arm.
- Colour light signal - a fixed lineside signal showing light indications to drivers.
- Cab signals - indications displayed in the driver's cab.

Certain signalling maintenance activities are partly dependent by the level of use of the network. The decision to renew signalling assets is regarded as not usage-related.

SLAB TRACK[3]

A form of railway track comprising a concrete base to which the chairs carrying the rails are secured. It eliminates the need for individual sleepers and ballast. The major advantages of the slab track are low maintenance, high availability and low weight. The installation of slab track is reported to cost about 20% more than ballasted track[2], but the maintenance costs can be reduced by 3 to 5 times that of ballasted track on a high-speed line (*e.g.* in Japan). If low levels of use are foreseen, or if low capital cost is a more important requirement, ballasted track would be the choice. For a heavily used railway, particularly one in a structurally restricted area like a tunnel or viaduct, non-ballasted track must be the best option on grounds of low maintenance cost and reduced space requirements.

SLEEPERS (or TIES)[3]

The sleepers are the transverse members of the track, made of wood, concrete or sometimes steel, which are used to secure the rails at the correct gauge. Steel chairs, fixed to the sleepers, hold the rails in place by means of clips or keys. Sleepers degrade in various ways, but under good conditions the wooden sleepers may last up to 25 years. Besides the environmental factors, such as corrosion, the effects of use also drive the costs of sleeper maintenance and renewal. The repeated impact by passing trains may cause fatigue, which leads, for example, to cracking or splitting and the failure of the rail fastening. The load imposed upon sleepers will vary with the track design, the quality of its maintenance, the quality of vehicle maintenance and the nature of the underlying track foundation.

SWITCHES (or POINTS)[3]

Switch is a track section, which allows the train to move from one track to another.

USAGE COST[1]

Usage cost is defined as that element of the total cost of maintenance and renewal, which varies with the amount and nature of traffic carried. Not all maintenance and renewal costs depend on usage. The costs of certain types of asset are considered not to vary at all and even where assets are considered to have usage related costs, the degree and significance of these costs varies by asset type.

TRACK[1]

The track is a fundamental part of the railway infrastructure. The usual track consists of the two steel rails, secured on sleepers to keep the rails at the correct distance apart and capable of supporting the weight of the trains. The third major component of the track is the ballast.

The costs of maintaining and renewing track assets are considered to be the most significant usage-related costs. These costs are also heavily influenced by the interaction between track and train. Poorly maintained vehicles impose greater track forces and hence cause greater track damage and more rapid degradation of the track assets; similarly, poorly maintained track imposes greater wear and thus higher rolling stock maintenance costs.

TRACK GEOMETRY[1]

Track geometry measurement means to measure the curve, grade and cross level condition of the track. To maintain the geometry of the track means to maintain its designed horizontal and vertical alignment and the dimension relationship between the two running rails. Track geometry maintenance can involve several processes, including manual attention. However, the most common form is mechanical tamping, in which special machines consolidate the ballast to support the corrected track geometry.

TRACK INFRASTRUCTURE

The track infrastructure consists of rails, sleepers, ballast, fastenings, switches, crossings and signals.

TRAFFIC DENSITY

Traffic density means the number of trains, which are running at the same time on a certain portion of a line.

TUNNEL

A structure provided to allow a railway line (or road) to pass under higher ground.

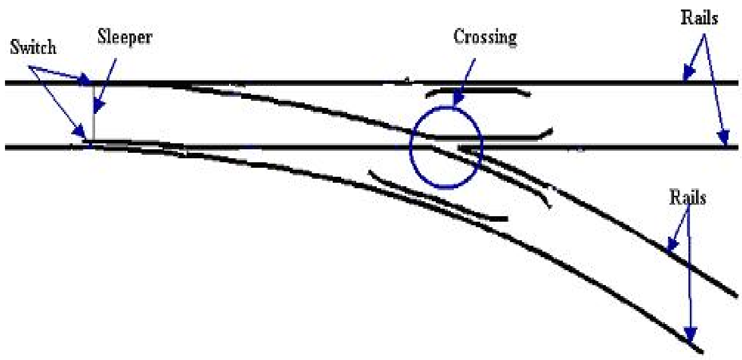

Figure A.2: Right-hand turnout

TURNOUT[2]

Turnouts are used to divide a track into two, three tracks. In the railway vocabulary, turnouts are referred to as "switch and crossing work". A turnout consists of a number of parts as follows: switches (see *Switches*), crossings (see *Crossings*) and rails (see *Rails*). On Figure A.2 a right-hand turnout is shown.

VISUAL INSPECTION[2]

The purpose of visual inspection is to check whether circumstances have arisen which may jeopardize safety of railway traffic. Inspection frequency varies depending on speed limit and daily tonnage from a few times a week on the most important lines to once a month on the least important lines. Extra inspections are necessary in exceptional circumstances, such as very hot weather.

Nederlandse Samenvatting (Summary in Dutch)

In de afgelopen jaren is veel aandacht besteed aan de noodzaak tot verbetering van de kwaliteit van Europese infrastructuurnetwerken voor verschillende wijzen van transport. In dit kader gaat speciale aandacht uit naar het openbaar vervoer in de belangrijke grootstedelijke gebieden.

Sinds geruime tijd is het in diverse landen regeringsbeleid om de congestie op de wegen te verminderen, aangezien dit ernstige economische en ecologische schade teweeg brengt. In Nederland neemt de regering bijvoorbeeld maatregelen om van het reizigers- en goederenvervoer per trein een volwaardig alternatief voor het wegvervoer te maken. Tot op zekere hoogte zijn deze maatregelen succesvol: in 2007 rapporteerden de Nederlandse Spoorwegen (NS) een stijging van 12% in personenkilometers ten opzichte van 2003 (NS homepage (2009)). Bovendien is ook het goederenvervoer per trein aanzienlijk toegenomen in de afgelopen jaren. Om te voldoen aan de toenemende vraag naar vervoer per spoor zullen meer treinen nodig zijn en ook meer frequente dienstverlening, met name in de spitsuren. In sommige gevallen is nieuwe infrastructuur nodig. Vanwege de toename van het aantal treinen en de verkeerslast neemt ook de slijtage van de infrastructuur toe. Daarom is meer onderhoud en vernieuwing nodig en dientengevolge neemt dus ook het tijdsbeslag toe van de onderhoudswerkzaamheden aan de infrastructuur. Vanwege de tijd dat de infrastructuur in beslag genomen wordt voor deze werkzaamheden zijn er minder beschikbare reismogelijkheden voor de treinen en kan mogelijk de infrastructuur niet voldoen aan de vraag.

Er zijn in de afgelopen jaren als gevolg van een aantal ernstige ongevallen strengere veiligheidsvoorschriften voor de werknemers aan het spoor ingevoerd. Zo is in sommige landen geen treinverkeer toegestaan tijdens onderhoudswerkzaamheden. Bovendien zijn de onderhoudskosten van de spoorweginfrastructuur aanzienlijk toegenomen in de afgelopen jaren. Bijvoorbeeld in Nederland stegen de onderhoudskosten met 40% tussen 1994 en 2001 en in 2006 bedroegen de kosten van het preventief onderhoud ongeveer 257 miljoen euro (Meeus and Staal (2007)).

Aangezien onderhoudswerkzaamheden aan de spoorweginfrastructuur zeer moeilijk zijn te plannen en het om hoge kosten gaat, is er een noodzaak voor de ontwikkeling van besliskundige methoden die de onderhoudsplanners helpen bij het komen tot een optimaal onderhoudsplan. De literatuurstudie in Hoofdstuk 3 laat zien dat er niet veel artikelen zijn die deze vraagstukken behandelen terwijl het gaat om een interessant onderwerp waarvan de resultaten kunnen bijdragen aan een verbetering van de kwaliteit van de spoorweginfrastructuur. Dit heeft ons ertoe aangezet om ons in dit proefschrift vooral op onderhoud van de spoorweginfrastructuur te concentreren. De wiskundige modellen en technieken ontwikkeld voor het plannen van het spoorwegenonderhoud kunnen ook worden toegepast voor onderhoudsplanning in andere publieke en/of private sectoren.

Een belangrijk facet dat in ogenschouw moet worden genomen bij het maken van een zorgvuldige onderhoudsplanning is hoe omgegaan kan worden met de wijzigingen in de dienstregeling van het treinverkeer, met name hoe de ergste hinder voor de treinreizigers en het personeel kan worden weggenomen. Dit houdt in, dat de dienstregeling en de materieelplanning bij het uitvoeren van onderhoudswerkzaamheden aangepast moeten worden, en wel op een zodanige manier dat het opnemen van de originele planning direct na afloop van de werkzaamheden geen probleem vormt. Bovenstaande problemen vormen de belangrijkste motivaties voor het proefschrift.

Dit proefschrift bestaat uit twee delen. In Deel I van dit proefschrift, dat bestaat uit de hoofdstukken 2, 3, 4 en 5, behandelen wij onderhoudsplanning in samenhang met productieplanning. Eerst beschouwen wij het probleem in het algemeen en vervolgens richten wij ons op toepassingen in de spoorwegen. In dit kader hebben we te maken met het probleem van het vinden van optimale tijdsintervallen voor het verrichten van routinematige onderhoudswerkzaamheden en grote projecten op een manier dat de kosten van de spoorbezetting en onderhoudskosten tot een minimum worden beperkt. In Deel II (Hoofdstuk 6) behandelen wij de noodzakelijke aanpassing van de materieelplanning voor het reizigersvervoer als gevolg van veranderende omstandigheden (bijvoorbeeld in verband met geplande onderhoudswerkzaamheden). Meer specifiek gaat het hier om het “Rolling Stock ReBalancing Problem”.

De belangrijkste onderzoeksdoelstellingen zijn als volgt geformuleerd:

1. Geef een overzicht van de bestaande literatuur over de relatie tussen onderhoudsplanning en productie.

2. Identificeer een aantal tactische en operationele problemen met betrekking tot onderhoudsplanning van de spoorweginfrastructuur en ontwikkel besliskundige modellen voor het ondersteunen van de besluitvorming. Onderzoek het effect van de planning van het onderhoud van de spoorweginfrastructuur op het treinvervoer en identificeer problemen bij de materieelplanning die optreden tijdens gepland onderhoud van de infrastructuur.

3. Analyseer de ontwikkelde modellen, onderzoek de rekenkundige complexiteit daarvan, ontwikkel oplossingsmethoden voor de modellen en test deze.

Hoofdbevindingen van dit proefschrift

Hieronder worden de belangrijkste resultaten van dit proefschrift per hoofdstuk vermeld.

In Hoofdstuk 2 hebben wij een overzicht gegeven van de literatuur over wiskundige modellen die de relatie tussen onderhoud en productie betreffen. Deze relatie bestaat op verschillende manieren. Allereerst is onderhoud bedoeld om de productie te faciliteren, maar voor het uitvoeren van onderhoud wordt de productie vaak gestopt. Om negatieve effecten van het stopzetten van de productie te elimineren (bijvoorbeeld hoge downtime kosten en productieverlies), moet bij de planning van onderhoud rekening worden gehouden met de behoefte om productie te leveren. Hier behandelen wij specifieke toepassingen van planningsmodellen in verschillende sectoren, zoals spoorwegen, snelwegen, luchtvaart en elektriciteitbedrijven. Ten tweede kan het onderhoud zelf ook worden gezien als een productieproces dat moet worden gepland. Tenslotte, omdat onderhoud productiecapaciteit wegneemt, moeten productie en onderhoud worden gepland op een gentegreerde manier. Hoewel veel artikelen zijn geschreven over de interactie tussen onderhoud en productie, is er nog steeds een gebrek aan studies over specifieke bedrijfstakken. Aangezien de vraag naar onderhoud toeneemt doordat publiek en bedrijven minder fouten en uitval tolereren en de kosten voor onderhoud ook waarschijnlijk toenemen, verwachten wij meer studies voor de ondersteuning van het onderhoud en de productie beslissingen.

In Hoofdstuk 3 gaan wij verder met het literatuuronderzoek over de onderhoudsplanning van de spoorweginfrastructuur en in het bijzonder over het effect van onderhoud op de capaciteit van het spoor. Het resulterende overzicht is gebaseerd op een algemene structurering van het proces van onderhoudsplanning in acht fasen. Op basis van deze literatuurstudie kunnen we concluderen dat het plannen van het preventieve onderhoudswerk op de spoorweginfrastructuur heel moeilijk is, omdat er veel randvoorwaarden in aanmerking moeten worden genomen. Ten eerste is er bij het uitvoeren van onderhoud aan het spoorinfrastructuur meestal sprake van veel storingen voor de reizigers (bijvoorbeeld vertragingen, geannuleerde treinen), terwijl het treinverkeer de lengte en de frequentie van de infrastructuurbezit beperkt. Bovendien zijn in de afgelopen jaren als gevolg van een aantal ernstige ongevallen de veiligheidsvoorschriften voor de onderhoudswerknemers aanmerkelijk aangescherpt. Daarnaast zijn de onderhoudskosten van de spoorweginfrastructuur aanzienlijk toegenomen in de afgelopen jaren. Dit alles toont aan dat er een noodzaak is voor het ontwikkelen van wiskundige modellen en technieken die de onderhoudsmanagers kunnen helpen bij het maken van goede (dat wil zeggen kosten-effectieve) schema's voor het uitvoeren van verschillende onderhoudsactiviteiten op de spoorweginfrastructuur.

In Hoofdstuk 4 introduceren wij het "Preventive Maintenance Scheduling Problem"(PMSP), waarin een schema voor de repeterende routinewerken en eenmalig projecten is te vinden zodanig dat de som van het bezits- en de onderhoudskosten wordt geminimaliseerd. De bezitskosten worden voornamelijk bepaald door de spoorbezettingstijd, dat wil zeggen de tijd dat een spoor in beslag wordt genomen voor onderhoud en niet kan worden gebruikt voor het treinverkeer. Om de bezitskosten te minimaliseren worden de onderhoudsactiviteiten zoveel mogelijk in dezelfde periode geclusterd. De bijdrage van dit hoofdstuk is tweeledig. Allereerst hebben wij een wiskundige formulering gedefinieerd van PMSP en van een extensie hiervan ("Restricted Preventive Maintenance Scheduling Problem"(RPMSP)) en hebben wij bewezen dat PMSP NP-moeilijk is. Ten tweede was het noodzakelijk om een aantal heuristische oplossingsmethoden te ontwikkelen die oplossingen genereren die de optimale oplossing dicht benaderen. De reden hiervoor was, dat het probleem van de onderhoudsplanning een complex optimalisatieprobleem is en voor een groot aantal testcases was het moeilijk en tijdrovend om het probleem optimaal op te lossen. In totaal zijn er vier heuristische methoden ontwikkeld. De tests toonden aan dat PMSP resulteerde in aanzienlijk lagere totale kosten die RPMSP, maar de berekeningstijd is veel langer en soms was het zelfs onmogelijk om een optimale oplossing te vinden binnen 3 uur. Bovendien bleken de heuristische methoden zeer snel te zijn (1–2 honderdsten van seconden), maar slechts twee heuristische methoden gaven aanvaardbare resultaten in vergelijking met de CPlex oplossingen. Toch was onze conclusie dat als de planners in zeer korte tijd een goede (bijna optimaal) schema willen hebben, RPMSP of een van de heuristische methoden effectief kunnen worden gebruikt.

In Hoofdstuk 5 gaan wij verder met de studie van de PMSP. Onze doelstelling was het ontwikkelen van betere technieken voor de oplossing van de PMSP dan de heuristische methoden van Hoofdstuk 4. In de literatuur zagen we dat met succes meta-heuristische algoritmen werden gebruikt voor het plannen van onderhoud door hun robuustheid en vermogen om snel te zoeken die bijdragen aan het verminderen van de rekenkundige complexiteit van grootschalige onderhoudsplanningproblemen. Op basis hiervan implementeerden we in Hoofdstuk 5 voor het oplossen van de PMSP genetische algoritmen, memetische algoritmen met drie lokale zoektechnieken (tabu search, simulated annealing en steepest hill climbing) en iteratieve heuristische methodes die dezelfde lokale zoekmethodes gebruikt als het memetische algoritme. Verder hebben we een "two-phase opportunities-based"heuristische methode ontwikkeld (GA_OPP), waarin in de eerste fase opportunities worden gegenereerd met behulp van een genetisch algoritme en in de tweede fase de preventieve werken op deze opportunities worden uitgevoerd. Van de berekende resultaten concluderen wij dat behalve de iteratieve heuristische methodes die zeer slechte resultaten geven, de andere oplossingsmethodes gemiddeld de oplossingen gevonden door de CPlex-solver verbeteren voor de gevallen met lage bezitskosten, terwijl voor de gevallen met hoge bezitskosten alleen de GA_OPP de CPlex oplossing kan verbeteren. Als wij de prestaties van de genetische algoritmen met de prestaties van de memetische algoritmen vergelijken, dan komen wij tot de conclusie dat de lokale zoektechnieken voor bijna alle gevallen tot een verbetering van de genetische algoritme leiden. Voor zover alleen het memetische algoritme in beschouwing wordt genomen, dan heeft het simulated annealing algoritme in het algemeen de beste resultaten, gevolgd door het de tabu search en tenslotte door de steepest hill climbing.

Als samenvatting van onze bevindingen voor onderhoudsplanningproblemen kunnen we zeggen dat de wiskundige modellen en technieken ontwikkeld in Deel I van dit proefschrift de onderhoudsmanagers kunnen helpen om in een redelijk korte tijd kosteneffectieve planningen te kunnen maken voor het uitvoeren van verschillende preventieve onderhoudsactiviteiten op de spoorweginfrastructuur. Bovendien zijn de modellen gepresenteerd in de hoofdstukken 4 en 5 slechts elementaire modellen, maar ze kunnen worden uitgebreid om veel verschillende praktische problemen op te lossen omdat er in werkelijkheid veel meer restricties zijn waarmee een onderhoudsplanner rekening moet houden. Tot slot zouden wij willen noemen, dat de wiskundige modellen en technieken ontwikkeld voor het plannen van spoorwegonderhoud ook kunnen worden toegepast voor onderhoudsplanning in andere publieke / private sectoren.

Hoofdstuk 6 is gericht op de herschikking van het materieel voor het reizigersvervoer als gevolg van veranderde omstandigheden. Meer specifiek beschouwen wij de "Rolling Stock Rebalancing Problem"(RSRP). Vanwege wijzigingen in het dienstregeling (bijvoorbeeld gepland onderhoud van de spoorinfrastructuur of ongeplande storingen), bevat de input materieel omloop voor een bepaalde planningsperiode een aantal off-balances. RSRP is het probleem om het input materieel omloop zodanig aan te passen dat het aantal niet opgeloste off-balances minimaal is, in combinatie met andere criteria die betrekking hebben op de efficiëntie, dienstverlening en robuustheid. Sinds RSRP op korte termijn en in de real-time planning fase ontstaat, is er behoefte aan snelle oplossingsmethoden. In Hoofdstuk 6 is er een heuristische methode ontwikkeld om RSRP op te lossen. De prestatie van deze heuristische methode wordt vergeleken met een heuristische methode van Maróti (2006) en met de prestaties van de exacte oplossing methode van Fioole *et al.* (2006) die wordt gebruikt bij de NS. De vergelijking van de resultaten wordt gedaan op basis van bepaalde (varianten van) real-life data van de NS. Uit de berekende resultaten hebben we geconcludeerd dat beide heuristische methoden zeer snel zijn en effectief kunnen worden gebruikt, niet alleen voor het oplossen van grote korte-termijn planning problemen, maar ook als basis voor het oplossen van real-time herschikkingsproblemen in het geval van een storing van het spoorwegsysteem.

Bibliography

Abdulwhab, A., R. Billinton, A. Eldamaty, and S. Faried (2004), Maintenance Scheduling Optimization Using a Genetic Algorithm (GA) with a Probabilistic Fitness Function, *Electric Power Components and Systems*, **32**, 1239–1254.

Al-Zubaidi, H. and A. Christer (1997), Maintenance Manpower Modelling for a Hospital Building Complex, *European Journal of Operational Research*, **99**, 603–618.

Alfieri, A., R. Groot, L. Kroon, and A. Schrijver (2006), Efficient Circulation of Railway Rolling Stock, *Transportation Science*, **40**, 378–391.

Ashayeri, J., A. Teelen, and W. Selen (1996), A production and maintenance planning model for the process industry, *International Journal of Production Research*, **34**, 3311–3326.

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. Makhdoum (1998), Integrated Production and Quality Model Under Various Preventive Maintenance Policies, *Journal of the Operational Research Society*, **49**, 840–853.

Ben-Daya, M. and M. Rahim (2001), Integrated Production, Quality & Maintenance Models: an Overview, in M. Rahim and M. Ben-Daya (eds.), *Integrated Models in Production Planning, Inventory, Quality, and Maintenance*, chap. 1, Kluwer Academic Publishers, pp. 3–28.

Ben-Khedher, N., J. Kintanar, C. Queille, and W. Stripling (1998), Schedule Optimization at SNCF: from Conception to Day of Departure, *Interfaces*, **28**, 6–23.

Bengü, G. (1994), Telecommunications Systems Maintenance, *Computers & Operations Research*, **21**, 337–351.

British Standard (1984), *Maintenance Management Terms in Terotechnology*, vol. BS:3811, BSI, London.

Brucker, P., J. Hurink, and T. Rolfes (2003), Routing of Railway Carriages, *Journal of Global Optimization*, **27**, 313–332.

Budai, G. and R. Dekker (2002), An Overview of Techniques Used in Planning Railway Infrastructure Maintenance, in W. Geraerds and D. Sherwin (eds.), *Proceedings of IFRIMmmm (maintenance management and modelling) conference, Växjö, Sweden*, Växjö University, Sweden.

Budai, G. and R. Dekker (2004), A Dynamic Approach for Planning Railway Maintenance Activities, in J. Allan, C. Brabbia, R. Hill, G. Sciutto, and S. Sone (eds.), *Computers in Railways IX*, WIT Press Southampton Boston.

Budai, G., R. Dekker, and U. Kaymak (2009b), Genetic and Memetic Algorithms for Scheduling Railway Maintenance Activities, *Computers and Operations Research*, submitted in February 2009.

Budai, G., R. Dekker, and R. Nicolai (2008), A Review of Planning Models for Maintenance & Production, in K. Kobbacy and D. Murthy (eds.), *Complex System Maintenance Handbook: Blending Theory with Practice*, chap. 13, Springer Verlag, London, pp. 321–344.

Budai, G., D. Huisman, and R. Dekker (2005), Scheduling of the Rolling Stock During Track Maintenance, in *Proceedings of IFORS Conference, Honolulu Hawaii*, INFORMS.

Budai, G., D. Huisman, and R. Dekker (2006), Scheduling Preventive Railway Maintenance Activities, *Journal of the Operational Research Society*, **57**, 1035–1044.

Budai, G., G. Maróti, R. Dekker, D. Huisman, and L. Kroon (2009a), Rescheduling in Passenger Railways: the Rolling Stock Rebalancing Problem, *Journal of Scheduling*, under second revision.

Burke, E., J. Clark, and A. Smith (1997a), Four Methods for Maintenance Scheduling, in *Proceeding of the International Conference on Artificial Neural Networks and Genetic Algorithms*, Springer.

Burke, E. and A. Smith (1997b), A Memetic Algorithm for the Maintenance Scheduling Problem, in *Proceedings of the ICONIP Conference, Dunedin, New Zealand*.

Burke, E. and A. Smith (1999a), A Memetic Algorithm to Schedule Planned Grid Maintenance, in *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation*, IOS Press.

Burke, E. and A. Smith (1999b), A Memetic Algorithm to Schedule Planned Maintenance for the National Grid, *The ACM Journal of Experimental Algorithmics*, **4**, 1–13.

Burke, E. and A. Smith (2000), Hybrid Evolutionary Techniques for the Maintenance Scheduling Problem, in *IEEE Transactions on Power Systems*, vol. 15.

Campbell, J. (1995), *Uptime: Strategies for Excellence in Maintenance Management*, Portland, Productivity Press.

Caprara, A., L. Kroon, M. Monaci, M. Peeters, and P. Toth (2007), Passenger Railway Optimization, in C. Barnhart and G. Laporte (eds.), *Handbook in OR & MS vol. 14*, Elsevier, pp. 129–187.

Carretero, J., J. Perez, F. Garcia-Carballeira, A. Calderon, J. Fernandez, J. Garcia, A. Lozano, L. Cardona, N. Cotaina, and P. Prete (2003), Applying RCM in Large Scale Systems: a Case Study with Railway Networks, *Reliability Engineering and System Safety*, **82**, 257–273.

Cassady, C., E. Pohl, and W. Murdock (2001), Selective Maintenance Modeling for Industrial Systems, *Journal of Quality in Maintenance Engineering*, **7**, 104–117.

Chan, W., T. Fwa, and K. Hoque (2001), Constraint Handling Methods in Pavement Maintenance Programming, *Transportation Research Part C*, **9**, 175–190.

Charles, A., I. Floru, C. Azzaro-Pantel, L. Pibouleau, and S. Domenech (2003), Optimization of Preventive Maintenance Strategies in a Multipurpose Batch Plant: Application to Semiconductor Manufacturing, *Computers and Chemical Engineering*, **27**, 449–467.

Chelbi, A. and D. Ait-Kadi (2004), Analysis of a Production/Inventory System with Randomly Failing Production Unit Submitted to Regular Preventive Maintenance, *European Journal of Operational Research*, **156**, 712–718.

Cheung, B., K. Chow, L. Hui, and A. Yong (1999), Railway Track Possession Assignment Using Constraint Satisfaction, *Engineering Applications of AI*, **12**, 599–611.

Cheung, K., C. Hui, H. Sakamoto, K. Hirata, and L. O'Young (2004), Short-term Sitewide Maintenance Scheduling, *Computers and Chemical Engineering*, **28**, 91–102.

Cho, D. and M. Parlar (1991), A Survey of Maintenance Models for Multi-unit Systems, *European Journal of Operational Research*, **51**, 1–23.

Chopra, S. and P. Meindl (2001), *Supply Chain Management: Strategy, Planning, and Operation*, Prentice Hall: New Jersey.

Christer, A. and W. Waller (1984), Delay Time Models of Industrial Maintenance Problems, *Journal of the Operational Research Society*, **35**, 401–406.

Chung, K. (2003), Approximations to Production Lot Sizing with Machine Breakdowns, *Computers & Operations Research*, **30**, 1499–1507.

Clausen, J., A. Larsen, and J. Larsen (2005), Disruption Management in the Airline Industry: Concepts, Models and Methods, Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, www2.imm.dtu.dk/pubdb/views/edoc_download.php/3763/pdf/imm3763.pdf (last accessed on 17 January 2009).

Cobb, R. (1995), Modeling Aircraft Repair Turntime: Simulation Supports Maintenance Marketing, *Journal of Air Transport Management*, **2**, 25–32.

Cohn, A. and C. Barnhart (2003), Improving Crew Scheduling by Incorporating Key Maintenance Routing Decisions, *Operations Research*, **51(3)**, 387–396.

Cross, M. (1988), Raising the Value of Maintenance in the Corporate Environment, *Management Research News*, **11**, 8–11.

Dagpunar, J. (1996), A Maintenance Model with Opportunities and Interrupt Replacement Options, *Journal of the Operational Research Society*, **47**, 1406–1409.

Dahal, K. and N. Chakpitak (2007), Generator Maintenance Scheduling in Power Systems Using Metaheuristic-based Hybrid Approaches, *Electric Power Sytems Research*, **77**, 771–779.

Das, T. and S. Sarkar (1999), Optimal Preventive Maintenance in a Production Inventory, *IIE Transactions*, **31**, 537–551.

De Ree, D. (2006), Scheduling Preventive Railway Maintenance Using Memetic Algorithms, Econometrics and Management Science, Erasmus University Rotterdam, Bachelor Thesis.

Dedopoulos, L. and N. Shah (1995), Preventive Maintenance Policy Optimisation for Multipurpose Plant Equipment, *Computers and Chemical Engineering*, **19**, 693–698.

Dekker, R. (1995), Integrating Optimisation, Priority Setting, Planning and Combining of Maintenance Activities, *European Journal of Operational Research*, **82**, 225–240.

Dekker, R. (1996), Applications of Maintenance Optimization Models: a Review and Analysis, *Reliability Engineering and System Safety*, **51**, 229–240.

Dekker, R. and M. Dijkstra (1992), Opportunity-based Age Replacement: Exponentially Distributed Times Between Opportunities, *Naval Research Logistics*, **39**, 175–190.

Dekker, R. and R. Plasmeijer (2001), Multi-parameter Maintenance Optimisation via the Marginal Cost Approach, *Journal of the Operational Research Society*, **52**, 188–197.

Dekker, R., R. Plasmeijer, and J. Swart (1998a), On the Economic Evaluation of the Joint-to-joint Concept for Road Maintenance, *IMA Journal of Mathematics Applied in Business and Industry*, **9**, 109–156.

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., F. van der Duyn Schouten, and R. Wildeman (1997), A Review of Multi-Component Maintenance Models, *Mathematical Methods of Operations Research*, **45**, 411–435.

Dekker, R., J. van der Meer, R. Plasmeijer, and R. Wildeman (1998b), Maintenance of Light-standards – a Case-Study, *Journal of the Operational Research Society*, **49**, 132–143.

Dekker, R., R. van Egmond, J. Frenk, and R. Wildeman (2000), A General Approach for the Coordination of Maintenance, in M. Ben-Daya, S. Duffuaa, and A. Raouf (eds.), *Maintenance, Modeling and Optimization*, Kluwer Academic Publishers: Boston/Dordrecht/London, pp. 245–282.

Dekker, R. and C. van Rijn (1996), *PROMPT - a Decision Support System for Opportunity Based Preventive Maintenance*, vol. 154, Springer-Verlag, Berlin, pp. 530–549.

Den Hertog, D., J. van Zante-de Fokkert, S. Sjamaar, and R. Beusmans (2005), Optimal Working Zone Division for Safe Track Maintenance in The Netherlands, *Accident Analysis and Prevention*, **37**, 890–893.

Dijkstra, M., L. Kroon, M. Salomon, J. van Nunen, and L. van Wassenhoven (1994), Planning the Size and Organization of KLM's Aircraft Maintenance Personnel, *Interfaces*, **24**, 47–58.

Edwards, D., G. Holt, and F. Harris (2002), Predicting Downtime Costs of Tracked Hydraulic Excavators Operating in the UK Opencast Mining Industry, *Construction Management & Economics*, **20**, 581–591.

El Moudani, W. and F. Mora-Camino (2000), A Dynamic Approach for Aircraft Assignment and Maintenance Scheduling by Airlines, *Journal of Air Transport Management*, **6**, 233–237.

Esveld, C. (2001), *Modern Railway Track*, MRT-Productions, Zaltbommel, The Netherlands.

Feo, T. and J. Bard (1989), Flight Scheduling and Maintenance Base Planning, *Management Science*, **35**, 1415–1432.

Ferreira, L. (1997b), Planning Australian Freight Rail Operations: an Overview, *Transportation Research - A*, **31**, 335–348.

Ferreira, L. and M. Murray (1997a), Modelling Rail Track Deterioration and Maintenance Current Practices and Future Needs, *Transport Reviews*, **17**, 207–221.

Finch, B. and J. Gilbert (1986), Developing Maintenance Craft Labor Efficiency Through an Integrated Planning and Control System: a Prescriptive Model, *Journal of Operations Management*, **6**, 449–459.

Fioole, P., L. Kroon, G. Maróti, and A. Schrijver (2006), A Rolling Stock Circulation Model for Combining and Splitting of Passenger Trains, *European Journal of Operational Research*, **174**, 1281–1297.

Freling, R., R. Lentink, L. Kroon, and D. Huisman (2005), Shunting of Passenger Train Units in a Railway Station, *Transportation Science*, **39**, 261–272.

Frost, D. and R. Dechter (1998), Optimizing with Constraints: a Case Study in Scheduling Maintenance of Electric Power Units, *Lecture Notes in Computer Science*, **1520**, 469–488.

Fwa, T., C. Tan, and W. Chan (1994), Road-maintenance Planning Using Genetic Algorithms II: Analysis, *Journal of Transportation Engineering, ASCE 120*, **5**, 710–722.

GAMS homepage (2009), World Wide Web page, http://www.gams.com (last accessed on 17 January 2009).

Garey, M. and D. Johnson (1979), *Computers and Intractability: a Guide to the Theory of NP-Completeness*, Freeman: San Francisco.

Geraerds, W. (1985), The Cost of Downtime for Maintenance: Preliminary Considerations, *Maintenance Management International*, **5**, 13–21.

Geraerds, W. (1991), The EUT Maintenance Model, in H. Bradley (ed.), *Operational Research 90, IFORS Conference, Athens, 1990*, Pergamon Press, New York.

Gharbi, A. and J. Kenne (2000), Production and Preventive Maintenance Rates Control for a Manufacturing System: an Experimental Design Approach, *International Journal of Production Economics*, **65**, 275–287.

Gharbi, A. and J. Kenne (2005), Maintenance Scheduling and Production Control of Multiple-machine Manufacturing Systems, *Computers & Industrial Engineering*, **48**, 693–707.

Gits, C. (1984), *On the Maintenance Concept for a Technical System: a Framework for Design*, Ph.D. thesis, Eindhoven University of Technology, The Netherlands.

Goel, H., J. Grievink, and M. Weijnen (2003), Integrated Optimal Reliable Design, Production, and Maintenance Planning for Multipurpose Process Plants, *Computers and Chemical Engineering*, **27**, 1543–1555.

Gopalan, R. and K. Talluri (1998), Mathematical Models in Airline Schedule Planning: a Survey, *Annals of Operations Research*, **76(1)**, 155–185.

Grimes, C. (1995), Application of Genetic Techniques to the Planning of Railway Track Maintenance Work, in A. Zalzala (ed.), *First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications, Sheffield, UK*, Conference Publication 414, IEE, University of Sheffield.

Groenevelt, H., L. Pintelon, and A. Seidmann (1992a), Production Lot Sizing with Machine Breakdowns, *Management Science*, **48**, 104–123.

Groenevelt, H., L. Pintelon, and A. Seidmann (1992b), Production Batching with Machine Breakdowns and Safety Stocks, *Operations Research*, **40**, 959–971.

Haghani, A. and Y. Shafahi (2002), Bus Maintenance Systems and Maintenance Scheduling: Model Formulations and Solutions, *Transportation Research Part A*, **36**, 453–482.

Hariga, M. (1994), A Deterministic Maintenance-Scheduling Problem for a Group of Non-identical Machines, *International Journal of Operations and Production Management*, **14**, 27–36.

Higgins, A. (1998), Scheduling of Railway Maintenance Activities and Crews, *Journal of the Operational Research Society*, **49**, 1026–1033.

Hokstad, P., H. Langseth, B. Lindqvist, and J. Vatn (2005), Failure Modeling and Maintenance Optimisation for a Railway Line, *International Journal of Performability Engineering.*, **1**, 51–64.

Holland, J. (1975), *Adaption in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor.

Huisman, D., L. Kroon, R. Lentink, and M. Vromans (2005), Operations Research in Passenger Railway Transportation, *Statistica Neerlandica*, **59**, 467–497.

ILOG homepage (2009), World Wide Web page, http://www.ilog.com (last accessed on 17 January 2009).

Improverail (2002), http://www.tis.pt/proj/improverail/Downloads/D6Final.pdf (last accessed on 30 November 2006).

Iravani, S. and I. Duenyas (2002), Integrated Maintenance and Production Control of a Deteriorating Production System, *IIE Transactions*, **34**, 423–435.

Jespersen Groth, J., D. Potthoff, J. Clausen, D. Huisman, L. Kroon, G. Maróti, and M. Nyhave Nielsen (2007), Disruption Management in Passenger Railway Transportation, Econometric Institute Report EI 2007-05, Erasmus University Rotterdam, The Netherlands.

Jovanovic, S. and I. Korpanec (2000), ECOTRACK: Experiences from Implementation and Use.

Jovanovic, S. and H. Zaalberg (2000), Ecotrack: Two Years of Experience, *Rail International*, 2–8.

Kelly, A. (1989), Maintenance and its Management, in *Conference Communications, Monks Hill, UK*.

Kenne, J. and E. Boukas (2003), Hierarchical Control of Production and Maintenance Rates in Manufacturing Systems, *Journal of Quality in Maintenance Engineering*, **9**, 66–82.

Kenne, J., E. Boukas, and A. Gharbi (2003), Control of Production and Corrective Maintenance Rates in a Multiple-machine, Multiple-product Manufacturing System, *Mathematical and Computer Modelling*, **38**, 351–365.

Kenne, J., A. Gharbi, and M. Beit (2007), Age-dependent Production Planning and Maintenance Strategies in Unreliable Manufacturing Systems with Lost Sale, *European Journal of Operational Research*, **178**, 408–420.

Kianfar, F. (2005), A Numerical Method to Approximate Optimal Production and Maintenance Plan in a Flexible Manufacturing System, *Applied Mathematics and Computation*, **170**, 924–940.

Knights, P., F. Jullian, and L. Jofre (2005), Assessing the “Size” of the Prize: Developing Business Cases for Maintenance Improvement Projects, in *Proceedings of the International Physical Asset Management Conference*.

Kobbacy, K. and D. Murthy (2007), An Overview, in K. Kobbacy and D. Murthy (eds.), *Complex System Maintenance Handbook: Blending Theory with Practice*, chap. 1, Springer Verlag, Berlin, pp. 3–18.

Kohl, N., A. Larsen, J. Larsen, A. Ross, and S. Tiourine (2007), Airline Disruption Management – Perspectives, Experiences and Outlook, *Journal of Air Transport Management*, **13**, 149–162.

Kralj, B. and R. Petrovic (1988), Optimal Preventive Maintenance Scheduling of Thermal Generating Units in Power Systems – A Survey of Problem Formulations and Solution Methods, *European Journal of Operational Research*, **35**, 1–15.

Kyriakidis, E. and T. Dimitrakos (2006), Optimal Preventive Maintenance of a Production System with an Intermediate Buffer, *European Journal of Operational Research*, **168**, 86–99.

Lam, K. and M. Rahim (2002), A Sensitivity Analysis of an Integrated Model for Joint Determination of Economoc Design of $\overline{x}$-control Charts, Economic Production Quantity and Production Run Length for a Deteriorating Production System, *Quality and Reliability Engineering International*, **18**, 305–320.

Lamson, S., N. Hastings, and R. Willis (1983), Minimum Cost Maintenance in Heavy Haul Rail Track, *Journal of the Operational Research Society*, **34**, 211–223.

Langdon, W. and P. Treleaven (1997), Scheduling Maintenance of Electrical Power Transmission Networks Using Genetic Programming, in K. Warwick, A. Ekwue, and R. Aggarwal (eds.), *Artificial Intelligence Techniques in Power Systems*, Institution of Electrical Engineers, Stevenage, UK.

Lapa, C., C. Pereira, and M. de Barros (2006), A Model for Preventive Maintenance Planning by Genetic Algorithms Based in Cost and Reliability, *Reliability Engineering and System Safety*, **91**, 233–240.

Lapa, C., C. Pereira, and A. Mol (2000), Maximization of a Nuclear System Availability Through Maintenance Scheduling Optimization Using a Genetic Algorithm, *Nuclear Engineering and Design*, **196**, 219–231.

Lee, H. (2005), A Cost/benefit Model for Investments in Inventory and Preventive Maintenance in an Imperfect Production System, *Computers and Industrial Engineering*, **48**, 55–68.

Lee, H. and M. Rosenblatt (1987), Simultaneous Determination of Production Cycle and Inspection Schedules in a Production System, *Management Science*, **33**, 1125–1137.

Lee, H. and M. Rosenblatt (1989), A Production and Maintenance Planning Model with Restoration Cost Dependent on Detection Delay, *IIE Transactions*, **21**, 368–375.

Lee, H. and M. Srinivasan (2001), A Production/inventory Policy for an Unreliable Machine, in M. Rahim and M. Ben-Daya (eds.), *Integrated Models in Production Planning, Inventory, Quality, and Maintenance*, Kluwer Academic Publishers.

Lentink, R. (2006), *Algorithmic Decision Support for Shunt Planning*, Ph.D. thesis, Erasmus University Rotterdam, The Netherlands.

Levi, D. (2001), Optimization of Track Renewal Policy, in *Proceedings of the 5th World Congress on Railway Research, Cologne, Germany*, www.sncf.com/wcrr/SESSIONS/467.PDF (last accessed on 25 August 2003).

Li, J., P. Mirchandani, and D. Borenstein (2007), Vehicle Rescheduling Problem: Model and Algorithms, *Networks*, **50**, 211–229.

Li, S., C. Ting, C. Lee, and S. Chen (2002), Maintenance Scheduling of Oil Storage Tanks Using Tabu-Based Genetic Algorithm, in *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence*.

Lin, G. and D. Gong (2006), On a Production-inventory System of Deteriorating Items Subject to Random Machine Breakdowns with a Fixed Repair Time, *Mathematics and Computer Modelling*, **43**, 920–932.

Lingaya, N., J. Cordeau, G. Desaulniers, J. Desrosiers, and F. Soumis (2002), Operational Car Assignment at VIA Rail Canada, *Transportation Research B*, **36**, 755–778.

Liu, C., A. Hammad, and Y. Itoh (1997), Maintenance Strategy Optimization of Bridge Decks Using Genetic Algorithm, *Journal of Transportation Engineering*, **123**, 91–100.

Makis, V. and J. Fung (1995), Optimal Preventive Replacement, Lot Sizing and Inspection Policy for a Deteriorating Production System, *Journal of Quality in Maintenance Engineering*, **1(4)**, 41–55.

Maróti, G. (2006), *Operations Research Models for Railway Rolling Stock Planning*, Ph.D. thesis, Eindhoven University of Technology, Eindhoven, The Netherlands.

Meeus, J. and H. Staal (2007), Ontspoord 2, *NRC Handelsblad*, 24 November 2007, http://www.nrc.nl/nieuwsthema/spoor/article1873678.ece/Ontspoord_deel_2 (in Dutch – last accessed on 3 January 2009).

Meier-Hirmer, C., F. Sourget, and M. Roussignol (2005), Optimising the Strategy of Track Maintenance, in K. Kolowrocki (ed.), *ESREL 2005. Advances in Safety and Reliability*, A.A. Balkema Publishers Leiden.

Miwa, M., T. Ishikawa, and T. Oyama (2001), Modeling the Optimal Decision-Making for Multiple Tie Tamper Operations, in *Proceedings of the 5th World Congress on Railway Research, Cologne, Germany*.

Morcous, G. and Z. Lounis (2005), Maintenance Optimization of Infrastructure Networks Using Genetic Algorithms, *Automation in Construction*, **14**, 129–142.

Moscato, P. and A. Schaerf (1997), Local Search Techniques for Scheduling Problems - A Tutorial, unpublished paper, www.cs.bgu.ac.il/ cp041/Andrea_Local_Search.ps (last accessed on 17 January 2009).

Munoz, A., S. Martorell, and V. Serradell (1997), Genetic Algorithms in Optimizing Surveillance and Maintenance of Components, *Reliability Engineering and System Safety*, **57**, 107–120.

Nahmias, S. (2005), *Production and Operations Analysis*, 5^{th} edn., McGraw-Hill, Boston.

Negnevitsky, M. (2005), *Artificial Intelligence*, Pearson Education, Essex.

Negnevitsky, M. and G. Kelareva (1999), Genetic Algorithms for Maintenance Scheduling in Power Systems, in *Australasian Universities Power Engineering Conference and IEAust Electric Energy Conference, Darwin*.

Nicolai, R., G. Budai, R. Dekker, and M. Vreijling (2004), Modelling the deterioration of the coating on steel structures: a comparison of methods, in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, Danvers, pp. 4177–4182.

NS homepage (2009), World Wide Web page, http://www.ns.nl (last accessed on 17 January 2009).

Office of the Rail Regulator (2001), The Possessions Review: Provisional Conclusions on the Schedule 4 Incentive Structure, http://www.rail-reg.gov.uk/upload/pdf/140-possess_main.pdf (last accessed on 14 november 2008).

Okamura, H., T. Dohi, and S. Osaki (2001), Computation Algorithms of Cost-effective EMQ Policies with PM, in M. Rahim and M. Ben-Daya (eds.), *Integrated Models in Production Planning, Inventory, Quality, and Maintenance*, Kluwer Academic Publishers.

Özekici, S. (1988), Optimal Periodic Replacement of Multi-Component Reliability Systems, *Operations Research*, **36**, 542–552.

Peeters, M. and L. Kroon (2008), Circulation of Railway Rolling Stock: a Branch-and-Price Approach, *Computers & Operations Research*, **35**, 538–556.

Pintelon, L., L. Gelders, and F. van Puyvelde (1997), *Maintenance Management*, Acco, Leuven.

Pistikopoulos, E., C. Vassiliadis, and L. Papageorgiou (2000), Process Design for Maintainability: an Optimization Approach, *Computers and Chemical Engineering*, **24**, 203–208.

Podofillini, L., E. Zio, and J. Vatn (2006), Risk-informed Optimisation of Railway Inspection and Maintenance Procedures, *Realibility Engineering and System Safety*, **91**, 20–35.

ProRail homepage (2009), World Wide Web page, http://www.prorail.nl (last accessed on 17 January 2009).

Radcliffe, N. and P. Surry (1994), Formal Memetic Algorithms, in T. Fogarty (ed.), *Evolutionary Computing: AISB Workshop*, Springer-Verlag.

Rahim, M. (1994), Joint Determination of Production Quantity, Inspection Schedule, and Control Chart Design, *IIE Transactions*, **26(6)**, 2–11.

Rahim, M. and M. Ben-Daya (1998), A Generalized Economic Model for Joint Determination of Production Run, Inspection Schedule and Control Chart Design, *International Journal of Production Research*, **36**, 277–289.

Rahim, M. and M. Ben-Daya (2001), Joint Determination of Production Quantity, Inspection Schedule, and Quality Control for an Imperfect Process with Deteriorating Products, *Journal of the Operational Research Society*, **52(12)**, 1370–1378.

Rosa, L. and B. Feiring (1995), Layout Problem for an Aircraft Maintenance Company Tool Room, *International Journal of Production Economics*, **40**, 219–230.

Rose, G. and D. Bennett (1992), Locating and Sizing Road Maintenance Depots, *European Journal of Operations Research*, **63**, 151–163.

Roundy, R. (1985), 98%-Effective Integer-Ratio Lot-Sizing for One-Warehouse Multi-Retailer Systems, *Management Science*, **31**, 1416–1430.

Sarper, H. (1993), Scheduling for the Maintenance of Completely Processed Low-demand Large Items, *Applied Mathematical Modelling*, **17**, 321–328.

Sastry, K., D. Goldberg, and G. Kendall (2005), Genetic Algorithms, in E. Burke and G. Kendall (eds.), *Search Methodologies - Introductory Tutorials in Optimization and Decision Support Techniques*, chap. 4, Springer US, pp. 97–125.

Savic, D., G. Walters, and J. Knezevic (1995a), Optimal Opportunistic Maintenance Policy Using Genetic Algorithms, 1: Formulation, *Journal of Quality in Maintenance Engineering*, **1**, 34–49.

Savic, D., G. Walters, and J. Knezevic (1995b), Optimal Opportunistic Maintenance Policy Using Genetic Algorithms, 2: Analysis, *Journal of Quality in Maintenance Engineering*, **1**, 25–34.

Shenoy, D. and B. Bhadury (1993), MRSRP–a Tool for Manpower Resources and Spares Requirements Planning, *Computers and Industrial Engineering*, **24**, 421–439.

Simson, S., L. Ferreira, and M. Murray (1999), Modelling Rail Track Maintenance Planning Using Simulation, in *Proceedings of The Australian Society for Operations Research 15th National Conference, Gold Coast, Brisbane*.

Srinivasan, M. and H. Lee (1996), Production-inventory Systems with Preventive Maintenance, *IIE Transactions*, **28**, 879–890.

Sriram, C. and A. Haghani (2003), An Optimization Model for Aircraft Maintenance Scheduling and Re-assignment, *Transportation Research Part A*, **37**, 29–48.

Sriskandarajah, C., A. Jardine, and C. Chan (1998), Maintenance Scheduling of Rolling Stock Using a Genetic Algorithm, *Journal of Operational Research Society*, **49**, 1130–1145.

Strukton Rail homepage (2009), World Wide Web page, http://www.struktonrail.nl (last accessed on 17 January 2009).

Swier, J. (2003), Outsourcing in The Netherlands; Experiences and Developments, Presentation, http://www.promain.org/images/brussel/ProRail_ProMain_okt_2003_2.pdf (last accessed on 28 November 2008).

Tagaras, G. (1988), An Integrated Cost Model for the Joint Optimization of Process Control and Maintenance, *Journal of the Operational Research Society*, **39(8)**, 757–766.

Tan, J. and M. Kramer (1997), A General Framework for Preventive Maintenance Optimization in Chemical Process Operations, *Computers and Chemical Engineering*, **21**, 1451–1469.

Tseng, S. (1996), Optimal Preventive Maintenance Policy for Deteriorating Production Systems, *IIE Transactions*, **28**, 687–694.

Van der Duyn Schouten, F., B. van Vlijmen, and S. Vos de Wael (1998), Replacement Policies for Traffic Control Signals, *IMA Journal of Mathematics Applied in Business and Industry*, **9**, 325–346.

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 Dijkhuizen, G. (2000), Maintenance Grouping in Multi-setup Multi-component Production Systems, in M. Ben-Daya, S. Duffuaa, and A. Raouf (eds.), *Maintenance, Modeling and Optimization*, Kluwer Academic Publishers.

Van Dijkhuizen, G. and A. van Harten (1997), Optimal Clustering of Frequency-constrained Maintenance Jobs with Shared Setups, *European Journal of Operational Research*, **99**, 552–568.

Van Zante-de Fokkert, J., D. den Hertog, F. van den Berg, and J. Verhoeven (2007), The Netherlands Schedules Track Maintenance to Improve Track Workers' Safety, *Interfaces*, **37**, 133–142.

Vatn, J., P. Hokstad, and L. Bodsberg (1996), An Overall Model for Maintenance Optimization, *Reliability Engineering and System Safety*, **51**, 241–257.

Vaurio, J. (1999), Availability and Cost Functions for Periodically Inspected Preventively Maintained Units, *Reliability Engineering and System Safety*, **63**, 133–140.

VolkerRail homepage (2009), World Wide Web page, http://www.volkerrail.nl (last accessed on 17 January 2009).

Wang, C. (2006), Optimal Production and Maintenance Policy for Imperfect Production Systems, *Naval Research Logistics*, **53**, 151–156.

Wang, C. and S. Sheu (2003), Determining the Optimal Production-maintenance Policy with Inspection Errors: Using a Markov Chain, *Computers & Operations Research*, **30**, 1–17.

Wang, H. (2002), A Survey of Maintenance Policies of Deteriorating Systems, *European Journal of Operational Research*, **139**, 469–489.

Weinstein, L. and C. Chung (1999), Integrating Maintenance and Production Decisions in a Hierarchical Production Planning Environment, *Computers & Operations Research*, **26**, 1059–1074.

Wijnmalen, D. and A. 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., R. Dekker, and A. Smit (1997), A Dynamic Policy for Grouping Maintenance Activities, *European Journal of Operational Research*, **99**, 530–551.

Wolsey, L. (1998), *Integer Programming*, New York, Wiley - Interscience.

Yan, S., T. Yang, and H. Chen (2004), Airline Short-term Maintenance Manpower Supply Planning, *Transportation Research Part A*, **38**, 615–642.

Yao, X., X. Xie, M. Fu, and S. Marcus (2005), Optimal Joint Preventive Maintenance and Production Policies, *Naval Research Logistics*, **52**, 668–681.

Zoeteman, A. (2001), A Life Cycle Costing Based Approach to Design and Maintenance of Railway Infrastructure - Providing Decision Support for Evaluating Track Designs of the Dutch High Speed Line.

Zoeteman, A. (2006), Asset Maintenance Management: State of the Art in the European Railways, *International Journal of Critical Infrastructures*, **2**, 171–186.

Zoeteman, A. (2007), ProRail's Management of Tracks and Turnouts, *European Railway Review*, http://www.europeanrailwayreview.com/articles/20080512_38 (last accessed on 24 November 2008).

Zoeteman, A. and E. Braaksma (2001), An Approach to Improving the Performance of Rail Systems in a Design Phase - A Practical Example from the High Speed Line South Tender, in *Proceedings of the 5th World Conference on Railway Research, Cologne*, http://pdf.ramssoft.de/IP_WCRR2001.pdf (last accessed on 30 September 2008).

Zoeteman, A. and R. van Zelm (2001), Developing a Cost-effective Renewal Strategy for Railway Track on Conventional Networks - A Practical Example of a Policy Revision at Netherlands Railways, in *Proceedings of the 5th World Conference on Railway Research, Cologne*, http://www.tbm.tudelft.nl/webstaf/arjenz/Publications.htm (last accessed on 30 November 2006).

Curriculum Vitae

Gabriella Budai–Balke was born on March 7th, 1975 in Reghin, Romania. After completing her secondary education in 1993 at the Petru Maior High School in Reghin, she commenced her higher education at the Faculty of Mathematics and Computer Science of the Babeş–Bolyai University in Cluj–Napoca, Romania. In 1997, Gabriella received her Bachelor's degree in Mathematics, upon which she taught mathematics at the Lucian Blaga High School in Reghin until 2000. In that year she also received a post–graduate degree in Computer Science at the Petru Maior University in Târgu–Mures, Romania. In 2000 Gabriella came to Rotterdam to participate in the MPhil programme in Economics at the Tinbergen Institute. She received her degree in 2002 with the specialisation in Operations Research, after which she became a Ph.D. candidate in the area of railway maintenance optimisation. In the course of her Ph.D. research Gabriella presented some of her research at international conferences. Moreover, one of her papers is published in the Journal of the Operational Research Society, another paper is published as a chapter in the Complex Systems Maintenance Handbook: Blending Theory with Practice and two other papers are submitted to Journal of Rescheduling and Computers & Operations Research. Gabriella has also been involved as teaching assistant and co-supervisor of bachelor's theses. Since 2007 Gabriella works as logistics consultant at Ortec B.V. in Gouda.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

406. S.S. FICCO, *Essays on imperfect information-processing in economics.*
407. P.J.P.M. VERSIJP, *Advances in the use of stochastic dominance in asset pricing.*
408. M.R. WILDENBEEST, *Consumer search and oligopolistic pricing: A theoretical and empirical inquiry.*
409. E. GUSTAFSSON-WRIGHT, *Baring the threads: Social capital, vulnerability and the well-being of children in Guatemala.*
410. S. YERGOU-WORKU, *Marriage markets and fertility in South Africa with comparisons to Britain and Sweden.*
411. J.F. SLIJKERMAN, *Financial stability in the EU.*
412. W.A. VAN DEN BERG, *Private equity acquisitions.*
413. Y. CHENG, *Selected topics on nonparametric conditional quantiles and risk theory.*
414. M.D. DE POOTER, *Modeling and forecasting stock return volatility and the term structure of interest rates.*
415. F. RAVAZZOLO, *Forecasting financial time series using model averaging.*
416. M.J.E. KABKI, *Transnationalism, local development and social security: the functioning of support networks in rural Ghana.*
417. M. POPLAWSKI RIBEIRO, *Fiscal policy under rules and restrictions.*
418. S.W. BISSESSUR, *Earnings, quality and earnings management: the role of accounting accruals.*
419. L. RATNOVSKI, *A Random Walk Down the Lombard Street: Essays on Banking.*
420. R.P. NICOLAI, *Maintenance models for systems subject to measurable deterioration.*
421. R.K. ANDADARI, *Local clusters in global value chains, a case study of wood furniture clusters in Central Java (Indonesia).*
422. V. KARTSEVA, *Designing Controls for Network Organizations: A Value-Based Approach.*
423. J. ARTS, *Essays on New Product Adoption and Diffusion.*
424. A. BABUS, *Essays on Networks: Theory and Applications.*
425. M. VAN DER VOORT, *Modelling Credit Derivatives.*
426. G. GARITA, *Financial Market Liberalization and Economic Growth.*
427. E. BEKKERS, *Essays on Firm Heterogeneity and Quality in International Trade.*
428. H. LEAHU, *Measure-Valued Differentiation for Finite Products of Measures: Theory and Applications.*
429. G. BALTUSSEN, *New Insights into Behavioral Finance.*

430. W. VERMEULEN, *Essays on Housing Supply, Land Use Regulation and Regional Labour Markets.*
431. I.S. BUHAI, *Essays on Labour Markets: Worker-Firm Dynamics, Occupational Segregation and Workplace Conditions.*
432. C. ZHOU, *On Extreme Value Statistics.*
433. M. VAN DER WEL, *Riskfree Rate Dynamics: Information, Trading, and State Space Modeling.*
434. S.M.W. PHLIPPEN, *Come Close and Co-Create: Proximities in pharmaceutical innovation networks.*
435. A.V.P.B. MONTEIRO, *The Dynamics of Corporate Credit Risk: An Intensity-based Econometric Analysis.*
436. S.T. TRAUTMANN, *Uncertainty in Individual and Social Decisions: Theory and Experiments.*
437. R. LORD, *Efficient pricing algorithms for exotic derivatives.*
438. R.P. WOLTHOFF, *Essays on Simultaneous Search Equilibrium.*
439. Y.-Y. TSENG, *Valuation of travel time reliability in passenger transport.*
440. M.C. NON, *Essays on Consumer Search and Interlocking Directorates.*
441. M. DE HAAN, *Family Background and Children's Schooling Outcomes.*
442. T. ZAVADIL, *Dynamic Econometric Analysis of Insurance Markets with Imperfect Information.*
443. I.A. MAZZA, *Essays on endogenous economic policy.*
444. R. HAIJEMA, *Solving large structured Markov Decision Problems for perishable-inventory management and traffic control.*
445. A.S.K. WONG, *Derivatives in Dynamic Markets.*
446. R. SEGERS, *Advances in Monitoring the Economy.*
447. F.M. VIEIDER, *Social Influences on Individual Decision Making Processes.*
448. L. PAN, *Poverty, Risk and Insurance: Evidence from Ethiopia and Yemen.*
449. B. TIEBEN, *The concept of equilibrium in different economic traditions: A Historical Investigation.*
450. P. HEEMEIJER, *Expectation Formation in Dynamic Market Experiments.*
451. A.S. BOOIJ, *Essays on the Measurement Sensitivity of Risk Aversion and Causal Effects in Education.*
452. M.I. LOPEZ YURDA, *Four Essays on Applied Microeconometrics.*
453. S. MEENTS, *The Influence of Sellers and the Intermediary on Buyers' Trust in C2C Electronic Marketplaces.*
454. S. VUJIC, *Econometric Studies to the Economic and Social Factors of Crime.*
455. F. HEUKELOM, *Kahneman and Tversky and the Making of Behavioral Economics.*