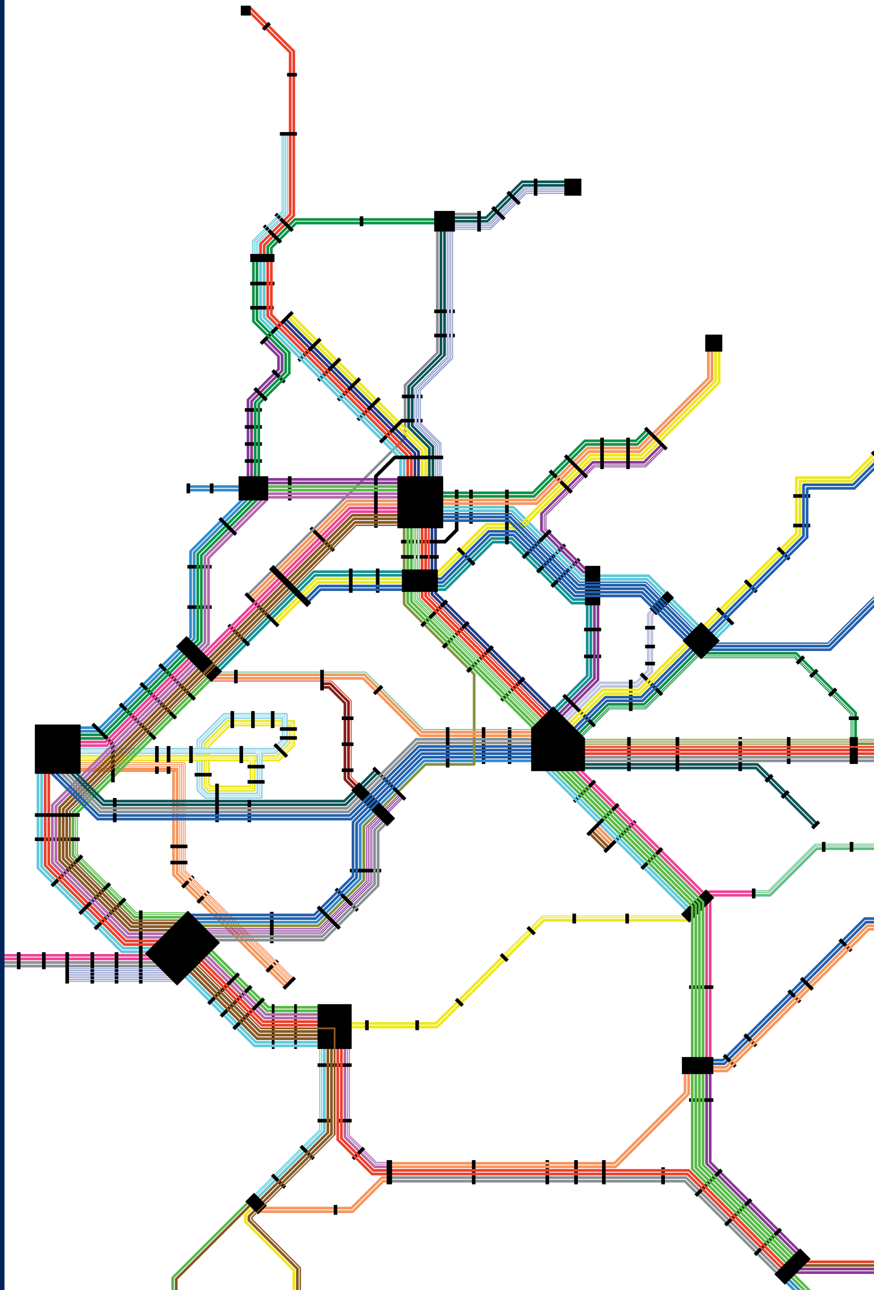
M.J.C.M. VROMANS

Reliability of Railway Systems



Reliability of Railway Systems

Michiel Vromans

Reliability of Railway Systems

Betrouwbaarheid van spoorwegsystemen

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus
Prof.dr. S.W.J. Lamberts
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 6 juli 2005 om 16.00 uur

door

Michiel Johannes Cornelius Maria Vromans

geboren te Veghel

Promotiecommissie

Promoteren:	Prof.dr. L.G. Kroon
	Prof.dr.ir. R. Dekker
	Prof.dr.ir. J.A.E.E. van Nunen
Overige leden:	Prof.dr. A.P.M. Wagelmans
	Prof.Dr.-Ing. I.A. Hansen
	Prof.Dr. A. Schöbel

Cover illustration: Part of the train line network in the Netherlands

Erasmus Research Institute of Management (ERIM)
Erasmus University Rotterdam
Internet: http://www.erim.nl
ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM Ph.D series Research in Management 62
TRAIL Thesis series T2005/7, The Netherlands TRAIL Research School

ISBN 90-5892-089-5

Acknowledgements

Writing this thesis has been a lot of hard work, but for most of the time it was a lot of fun as well. Railways are not only the topic of my thesis, but have also been a hot topic in the media: train personnel not liking their schedules, much too expensive new railway lines, overcrowded trains, and above all *delays*. When studying the reliability of railway systems, there is always something to talk about at parties.

Despite the fact that one doesn't finish his or her thesis during social talks, there is no chance of finishing the job without other people. In fact, the summary below is just an attempt to cover the people who have really been indispensable for completing my dissertation.

First of all, I would like to thank my promotor Leo Kroon, without whom I'd probably never even started writing a thesis. As my internship supervisor for my Master's degree, he introduced me to the fascinating railway world. Although the crew duties that I created led to the detested *Rondjes om de kerk* and high tensions within Netherlands Railways, Leo asked me to apply for some more research. Not knowing the hornets' nest ahead, I greedily accepted the opportunity to do research on railway delays. He also provided me with useful tips and innovative ideas. Patiently he read this thesis over-and-over again, before it could be published.

Rommert Dekker was my second promotor. He was always able to quickly understand the problems at hand. And with his great knowledge of a wide range of logistic processes, he guided me through the labyrinth named science. A further advantage for me was his drive to send me to conferences far away.

Jo van Nunen was my mental coach. Even in difficult circumstances, he could push me forward. His presence prevented me from deviating too far from the ideal path, and his overall thesis management made a five year time-line feasible.

All the other colleagues at the Erasmus should also be mentioned here. Although I have been in Utrecht more often than at the university, my fellow researchers, especially my fellow-lunchers, have made my period at the university an enjoyable one. My first roommate and also railway Ph.D.-candidate Leon Peeters has introduced me to the practical ins-and-outs of doing research and writing in LaTeX. When Leon left for Germany and Switzerland, the other fellow railway Ph.D.-candidate Ramon Lentink took over the LaTeX-part. Besides that, Ramon's presence has always had a positive influence on my output. I hope, and I'm convinced, he will finish his thesis shortly. For the last year, Niels Lang was my roommate. Where I was simulating trains, he was simulating ships: my trains were faster, but his ships carried more cookies.

Thanks to the work of Leo Kroon, Rommert Dekker, Matteo Fischetti, Erwin Abbink, Dennis Huisman, Ramon Lentink and Gerrit Timmer several co-authored papers are published, or will be published shortly.

The great working atmosphere at the logistics department of NS (Netherlands Railways) has always pushed me forward. Having moved four or five times, and because of the transitive nature of intern-students, a great variety of roommates has passed by. Their Belgian or Berghem-dialect, their Hungarian curses, stories and pictures of trips around the world, their presence at Stairways, and their use of Helpdesk2, all made it a lively and fruitful environment. The cooperation with ProRail has also been useful, particularly with respect to the simulation part of this thesis. Additionally, my new colleagues have helped me to survive the last six months.

Besides my promotors, I want to thank Cees Boxum, Chris Coppens, Michiel van Dellen, Monique Geerdink, Gerbrand Hulshof, Ramon Lentink and Marger Verschuur for reading and correcting smaller or larger parts of my thesis. Often overwhelmed by non-understandable mathematics, they critically reviewed my English.

Maar ook vrienden en familie die zich niet aan mijn Engels gewaagd hebben, hebben een belangrijke rol gespeeld. Dat zijn er te veel om op te noemen, maar bovenal wil ik mijn vader en moeder bedanken, die me al ruim 29 jaar met raad en daad bijstaan. Ook de afgelopen jaren hebben ze altijd achter me gestaan en me door de moeilijke momenten heen gesleept. Pap en mam: bedankt!

Michiel Vromans
Rotterdam, May 2005

Contents

Chapter 1

Introduction

The growing mobility in the Netherlands puts pressure on both the road and the railway network (V&W, 2000). High quality railway services are needed to facilitate the increasing numbers of passengers and goods. This is necessary to relieve the already congested roads. Especially in the crowded western part of the Netherlands, public transport plays a key role in the accessibility of urban areas. This is one of the key factors for economic growth. The same situation can be found in other urban areas throughout Europe.

The high costs of railway infrastructure ask for an increased utilization of the existing infrastructure. Since the Netherlands have one of the busiest railway networks in the world (Poort, 2002), this network is very vulnerable to disturbances. Keeping everything else the same, an even higher capacity consumption would only lead to an even larger vulnerability, whereas it is well known from queueing theory that average waiting times increase rapidly when a stochastic system approaches saturation (Tijms, 2003). Furthermore, a higher capacity consumption of the network also increases the snowball effect of delays.

Attractive railway services can only be offered with more reliable rolling stock and a more reliable infrastructure. However, to keep a high quality standard of operations, smarter methods of timetable construction are indispensable, since existing methods have major shortcomings. This thesis is a first step in that direction: this thesis discusses the improvement of railway reliability. More specifically, it studies the development of tools and guidelines which are helpful in the construction of reliable timetables.

1.1 Railway Developments in the Netherlands

1.1.1 The Dutch Railway Industry

In 1988 the Dutch government published the report "Second scheme for the structure of traffic and transport" (V&W, 1988) to give its view on traffic and transport in the long run. Using this report as a reference point, Netherlands Railways (in Dutch *Nederlandse Spoorwegen* or *NS*) started the project "Rail 21" that same year (NS, 1988, 1990, 1992). At that time, NS were more or less the only player in the Dutch railway market. It owned and maintained the infrastructure, and it operated both the cargo and passenger trains.

Developing their railway policy, the government and NS had to work in accordance with a directive of the European Union (EC, 1991). This directive states that the operations of railway traffic and the management of railway infrastructure should be separated. The impact of the directive on the Dutch situation was studied by the commission Wijffels and the commission Brokx (Wijffels et al., 1992, Commission Brokx, 1995), and by NS themselves (NS, 1994).

The government decided that NS had to be divided into a train operating and an infrastructure managing part. Moreover, the operating part should gain independence. Furthermore, competition was supported in an attempt to bring more efficiency into public transport. This led to several operators on one hand, and governmental infrastructure managers on the other hand.

The old NS from before this process have been divided into five parts: a reduced NS operating under the old name, Railion, Rail Traffic Control, Rail Infrastructure Management and Railned. The reduced NS are only responsible for operating the passenger trains on the Dutch network. Although it is officially an independent company since 1995, all the shares are still owned by the Dutch government. The former NS Cargo, responsible for freight trains, has merged with Deutsche Bahn Cargo under the new name Railion. Rail Traffic Control is responsible for operating the signals and switches, and for dispatching in case of disturbances. Rail Infrastructure Management is responsible for building new, and maintaining existing infrastructures. In addition to the tracks, this also includes the catenary and the signaling system. Finally, Railned is what could be called the railway capacity coordinator or railway referee. For each operating year, lasting from December to the next December, Railned allocates the network capacity among the operators. Furthermore, Railned is responsible for safety and it performs long term capacity research. These three organizations, Rail Traffic Control, Rail Infrastructure Management, and Railned, are directly working for and on account of the government. Since 2003 they operate under the common name ProRail, and they officially merged in 2005. The railway safety section was transferred to the Ministry of Transport.

Besides NS and Railion, several other operators have entered the Dutch railway market. Figure 1.1 shows the Dutch railway network and its passenger operators. At present, four passenger operators are active, of which NS are still by far the largest. NoordNed operates a few lines in the northern part of the country, Syntus is active in the east, and Connexxion operates another small line in the east. These lines in the north and the east have been tendered and won by the respective operator. At this moment and in the near future, many more peripheral lines, now operated by NS, have to be tendered. NS will bid on most of these lines as well. Until 2015, NS have the exclusive right of operating passenger trains on the *core network*, which is the backbone of the Dutch passenger train services. Additionally, NS offer international services in cooperation with foreign railway operators.

Unlike the passenger traffic, cargo operators do not have their own lines. In theory they are allowed to operate on the whole network. Railion, which includes the former NS Cargo, is still the largest cargo operator. The most important other freight operators are ACTS, Rail4Chem and ERS.

After its division, the reduced NS radically increased the number of passenger trains in the western and central parts of the Netherlands in 1996. Aided by a booming economy this led to a large increase in the number of passengers. However, the increased capacity consumption of the tracks is at least partially the reason for a decline in punctuality in 1996 and 1997 (see Figure 1.2, and Section 1.2 for the definition of punctuality). However, during the next few years the punctuality was on the rise again until mid 2000.

Then NS experienced a sudden, but rapid decline in punctuality. This decline was due to a shortage of rolling stock and personnel, newly structured crew schedules disliked by the personnel (see Nordbeck, 2003), bad communication between and within railway organizations, and a deteriorating state of the infrastructure. In January 2002, the CEO of NS had to resign, due to a punctuality figure of 79.9%, just short of the 80% agreed upon with the Ministry of Transport.

New incoming rolling stock, newly recruited train personnel, an agreement about new crew schedules, improved communication, and large investments in the infrastructure have all helped to improve the railway reliability. The 3-minute arrival punctuality was 86.0% over 2004.

Meanwhile, in 2000, the Dutch government published its view on transport, including railway transport, in the National Traffic and Transport Plan (V&W, 2000), which is a result of the “Second scheme for the structure of traffic and transport”, the

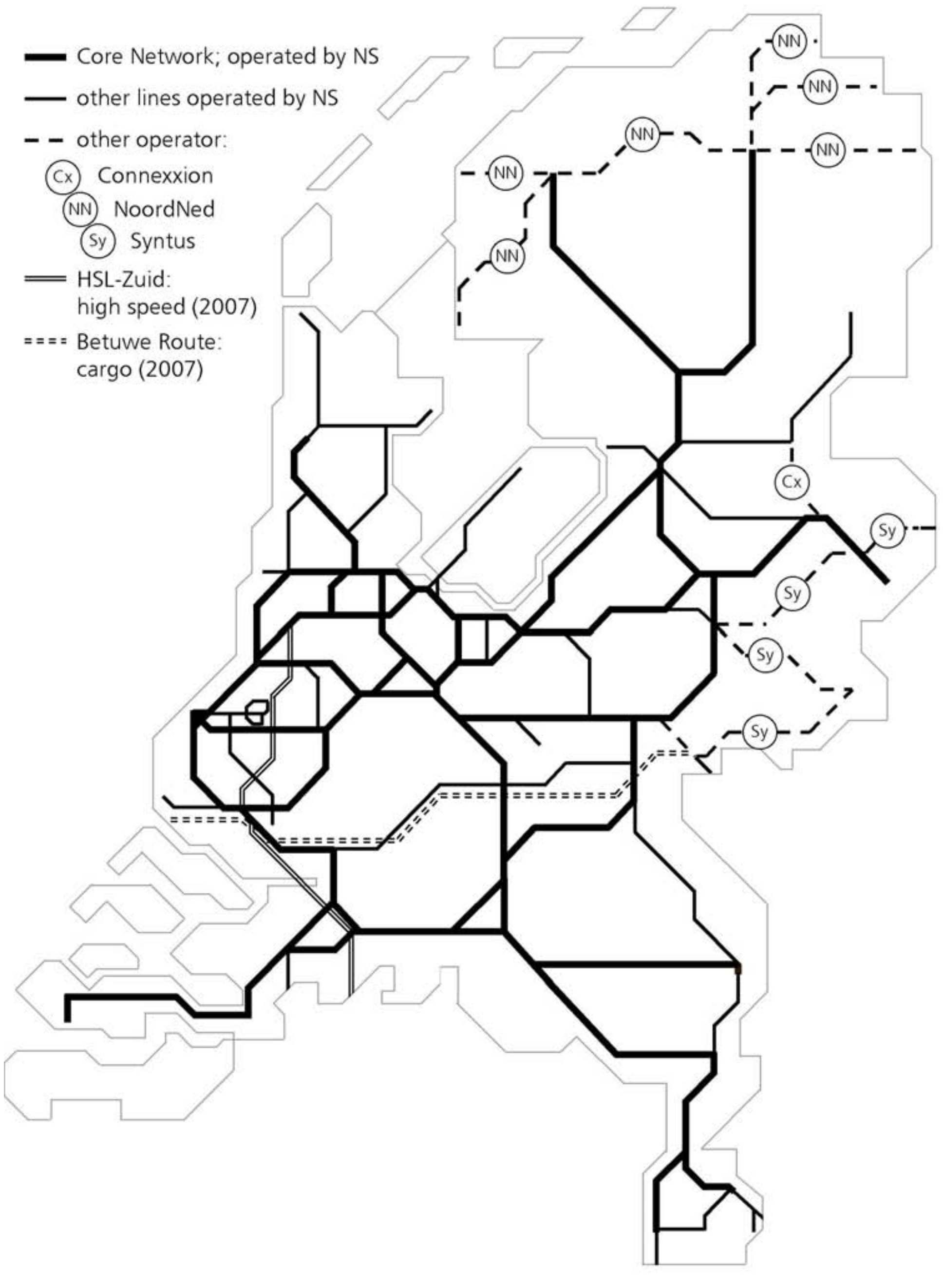


Figure 1.1: Dutch network and passenger operators (2004).

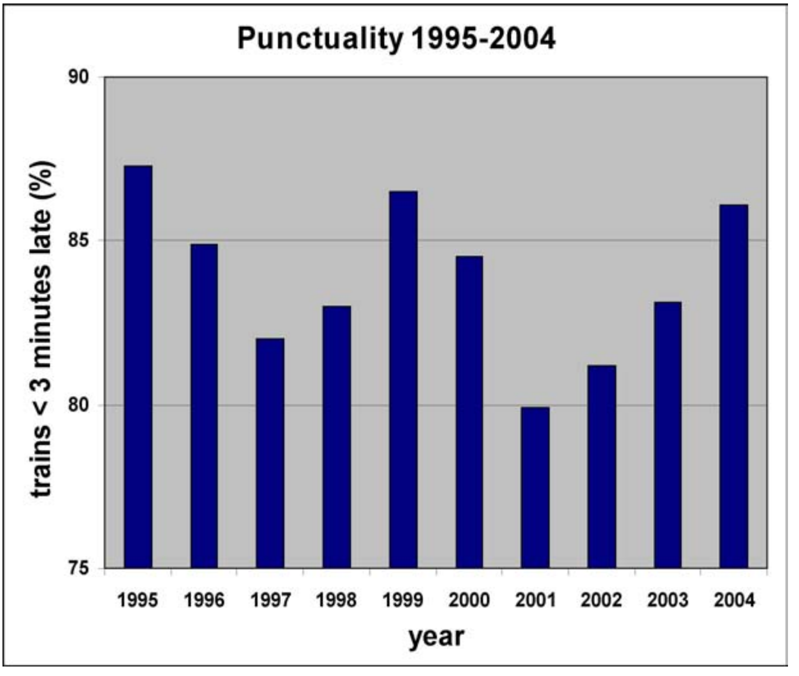


Figure 1.2: Punctuality in the Netherlands: 1995-2004. RailVerkeersleiding, 2004

EC-directive, privatization, competition, and other developments during the 1990's. It envisages a demand increase of 70% (and 100% in the west) for passenger railway transport, and an increase of 100% for cargo transport until 2020.

This forecasted increase, together with the punctuality problems, forced all parties in the railway sector to cooperate and to work out a plan. Without intelligent solutions it is impossible to facilitate a 70% to 100% demand increase on the already crowded Dutch railway network. The Dutch Ministry of Transport, the governmental organization ProRail, the passenger operator NS, and the cargo operators Railion, ACTS and Rail4Chem now work together in the project "Utilize and Build". In this project, the participants search for intelligent ways to increase the utilization and to improve the reliability at the same time. A summary of their ideas can be found in NS et al. (2003).

In 2007, several important infrastructure extensions will be finished. The high speed line from Amsterdam (in fact Schiphol Airport) via Rotterdam to Brussels and Paris (HSL-South, see Figure 1.1) will be ready. This line is for passenger traffic only, and will be operated by the High Speed Alliance (HSA), a joint-venture of NS (90%) and Air France-KLM (10%). A new cargo line, the Betuwe Route (see Figure 1.1), will connect the port of Rotterdam with its German hinterland. The cargo on this line will probably be transported by several operators. A final important project to be finished then, is the doubling of the line Amsterdam–Utrecht from 2 to 4 tracks. As part of the core network, passenger services will be operated by NS on this line. Also the cargo services will continue to be operated on this line. Where a new yearly passenger timetable usually shows a large resemblance with the one of the year before, a complete redesign is expected for the 2007 timetable.

1.1.2 Structure of Netherlands Railways

NS are by far the largest passenger operator in the Netherlands. Because of the particular focus in this thesis on Dutch passenger trains, NS timetables were used for the practical cases worked out in later chapters. A short overview of NS is given here to better understand their role and that of others in reliability improvement.

Officially, NS were transferred into an independent company in 1995. However, 100% of their shares are still owned by the government, but the government has little direct influence on the policies of the company. This is due to the fact that shareholders can not dismiss the Executive Board. However, there are still many financial and political ties between the government and NS, such as the *Performance Contract* (see later in this section).

The main responsibility of NS involves the passenger railway services on most of the Dutch railway network. As described above, most other tasks and responsibilities have been transferred to other organizations.

To offer high quality railway services for its clients, NS defined five main goals: (*i*) running on time, (*ii*) providing information and service, (*iii*) contributing to personal safety, (*iv*) creating sufficient transport capacity, and (*v*) offering clean trains and stations (NS, 2004a).

The organizational structure of NS in 2004 can be found in Figure 1.3. The NS organization is supported by ten corporate departments, such as Finance and Administration, and Corporate Communication, which are part of the NS Holding. The six dark colored organizations at the bottom of the figure form the core of the company. A short outline of these business units is given below. Furthermore there are three supporting companies: NS Education, NS Insurance, and NS Personnel Administration. Finally, NS has many subsidiaries. The HSA, for operating the Dutch high speed line, was already mentioned. Another important subsidiary is Strukton, a railway construction company. Below we briefly list the activities of each business unit.

Figure 1.3: Organization structure of NS (NS (Nederlandse Spoorwegen), 2004c).

- **NS Commerce** NS Commerce translates the vision of NS and the desires of the passengers into coherent designs and specifications of products. This should lead to a better service for the clients.

- **NS International** NS International runs, in cooperation with foreign railway companies, trains to Belgium, France, England, Germany and Switzerland.

- **NedTrain** NedTrain is responsible for the availability, safety and quality of the rolling stock. NedTrain cleans the train units, refuels them if necessary, is responsible for the maintenance, and sometimes refurbishes the rolling stock.

- **NS Passengers** The core business of NS is transporting passengers from A to B. Having said that, NS Passengers (in Dutch *NS Reizigers* or *NSR*), is without doubt the core of the company. It is responsible for offering train services to its clients. After planning the timetable, the rolling stock and train crew, it is also responsible for operating these plans.

- **NS Stations** NS Stations manages everything at and around stations. It makes sure that the stations are clean and maintained, it is responsible for ticket sales and services to the passengers, and it exploits the shops and restaurants in the stations. Sometimes they are also responsible for parking lots and other areas around a station.

- **NS Real Estate** NS own much of the real estate at and around stations. NS Real Estate invests in these areas, and develops attractive new business locations, all at top locations near public transport nodes.

Each day, NS operate about 5,000 passenger trains on a network of 2,800 km for close to 325,000 train kilometers. There are about 1,000,000 passenger journeys each day, with an average length of 44 km. The crisscrossed network of 100 different train lines along the 380 stations makes sure that almost 80% of the passenger trips is made without transfer. The 2,750 passenger coaches have about 225,000 seats available.

NS have over 24,800 employees and a total turnover of €2.7 billion in 2004. NS as a whole earned a profit of €144 million. The passenger services of NSR earned a profit for the first time in several years (€49 million). The remainder of the profit is mainly generated by NS Stations and NS Real Estate.

Contracts The so-called *Performance Contract* between the Ministry of Transport and NS gives NS the exclusive right of operating passenger trains on the *core network* until 2015. Figure 1.1 shows this core network. In return, NS have to offer a minimal service frequency, and they have to achieve a certain percentage of on-time arrivals of their passenger trains. NS will incur a fine if this percentage is not achieved. On the other hand, they will receive a premium when they can increase the number of passengers during the morning rush hours. Furthermore, upper limits on ticket prices are defined in the contract.

In order to compensate passengers for the mediocre performance by then, NS did not increase its ticket prices on January 1, 2002. Because of improved performance over that same year, NS wanted to implement the 2002 increase on December 31, 2002, followed by the regular increase of 2003 on the next day. Due to protests of the passenger organizations, this was forbidden in court. However, NS reached an agreement with the same passenger organizations: the *Agreement on Price Increase*. This agreement allows NS to increase its prices (on top of the regular increase) when the punctuality reaches a certain level. Here punctuality is the percentage of trains arriving at a major station with a delay less than three minutes. In fact, a 2.075% increase was allowed, and was already implemented in July 2004, when a punctuality of 84.4% over a twelve month period was reached. A second increase of 2.075% is allowed when NS reach a punctuality of 86.8%.

Both the Performance Contract and the Agreement on Price Increase contain explicit punctuality figures for NS. By reaching those punctualities, NS avoid penalties and further price increases are allowed. Additionally, in December 2004, the Dutch parliament has summoned the Minister of Transport to formulate stricter punctuality goals for NS. This stresses the importance of punctuality.

1.2 Reliability, Robustness, Stability, and Punctuality

The practical railway environment for this research was sketched in Section 1.1. In this section a few more theoretical concepts are discussed, which are important for understanding the thesis. Most of the thesis is built around these concepts. The first concept is *reliability*.

Reliability Reliability is the ability of a system or component to perform its required functions under stated conditions for a specified period of time (IEEE 1990).

Often used measures for reliability are the Mean Time Between Failures and the percentage of the planned processes that is completed in time. However, a wide range of definitions can be found, mainly depending on the system or application. Google (2005) already provides about 25 definitions.

In this thesis, the term reliability is used quite broadly. When a railway system is *reliable*, the trains run properly most of the time. This means that most of the passengers and goods are transported at the scheduled time. Only a small portion of the trains has delays or is not operated at all. Both the average delay and the variation in the delays are low.

Although reliability is obviously an important characteristic of any transportation system, it is difficult to give an unambiguous quantification for railways. Rietveld et al. (2001) provide the reader with seven possible measures:

1. punctuality, i.e. the probability that a train arrives less than x minutes late;
2. the probability of an early departure;
3. the mean difference between the expected arrival time and the scheduled arrival time;
4. the mean delay of an arrival given that one arrives late;
5. the mean delay of an arrival given that one arrives more than x minutes late;
6. the standard deviation of arrival times;
7. the adjusted standard deviation of the arrival times (ignoring the early arrivals), and various other more complex measures to represent the seriousness of unreliability.

However, many more measures for reliability exist. Our reliability objective in most of this thesis is similar to point (3): minimize the positive difference between the arrival times in operations and the scheduled arrival times. Minimizing this aspect has in many cases beneficial effects on all others.

Two other concepts are highly correlated to *reliability* and need to be introduced as well, viz. *robustness* and *stability*.

Robustness The robustness of a railway system indicates the influenceability of the system by *disturbances* (see below). A robust railway system can function fairly well under difficult circumstances. When a railway system is not robust, small external influences cause large delays which propagate quickly throughout the system in place and time.

Stability The formal engineering definition of stability is that a (repetitive) system is stable if it is able to perform its planned jobs within the planned time under undisturbed circumstances. The stability of a railway system is interpreted broader here, because we are also interested in the system's ability to return to normal circumstances. Therefore, stability is a measure for the time and effort which are needed to return to normal operations after a disturbance. A disturbed situation can, in a stable environment, return to normal operations quickly. When a system is instable, traffic will be irregular for a long time.

Also the following concepts are relevant:

Arrival and departure delays A delay is the positive difference between the planned time of an event and its actual realization time. In railway systems an event is a departure, an arrival or a passing time.

Disturbances Disturbances are mistakes, malfunctions or deviating conditions within a railway system or its environment, which can influence the railway traffic. Disturbances have many different causes, for which many different railway organizations are responsible. Moreover, a sizable portion of the disturbances has an external cause. For these external causes it is difficult and expensive to reduce their occurrence. Section 2.4 discusses disturbances in more detail.

Primary delays Primary delays, also called initial delays or source delays, are those delays which are not caused by other delayed trains. They are caused by a disturbance as described above. Note that slack in the timetable can reduce the size of a disturbance before it is measured as primary delay.

Secondary delays Secondary delays, or knock-on delays, are delays which are caused by earlier delays. Due to the interdependencies in railway systems, which are described in Section 2.5, a large part of the delays consists of secondary delays. The crowdedness of the Dutch railway network induces strong dependencies between train services. However, to date there is little quantitative insight in these dependencies. The dependencies are created by the timetable and other logistic plans, such as the rolling stock circulation. Because most railway experts in the Netherlands believe that there are more secondary than primary delays, it is worthwhile to look into them. Costs to reduce the disturbances and primary delays are usually high. Therefore this thesis aims at finding possibilities to reduce the dependencies within the railway plans, and the consequent secondary delays, at much lower costs.

Note that passengers do not see the difference between primary and secondary delays. For them a delay is a delay.

Punctuality One of the most used performance measures in railway systems is punctuality. In railway systems this is the percentage of trains arriving within a certain margin from the scheduled arrival time. Several official contracts between NS on one hand, and the government or passenger organizations on the other hand, use the 3-minute arrival punctuality as the key performance indicator. Other performance indicators are discussed in Section 2.7.

Despite the perception of the Dutch public and government, the on-time performance of trains is relatively high in the Netherlands. Figure 1.4 shows a punctuality comparison with other European countries from 2000. This is with a 5-minute margin, as is common in international comparisons. We see that, by then, Switzerland was the champion, but the Netherlands were number two, just ahead of Germany, Denmark, Belgium and France.

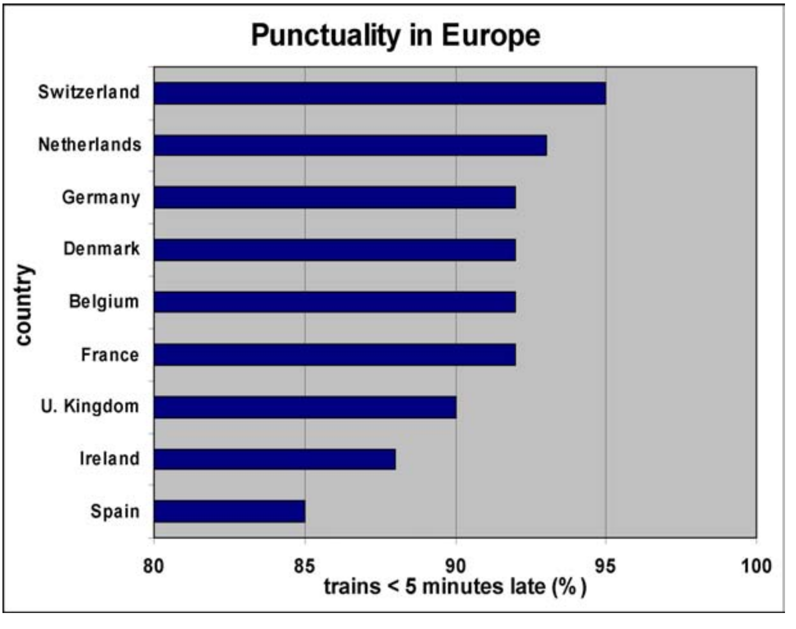


Figure 1.4: Punctuality in Europe (5 minutes; NS (2000)).

Network utilization This relatively high punctuality is attained despite the high utilization of the Dutch railway network. Here, the utilization is measured as the number of train kilometers per network kilometer per year. Figure 1.5 shows the network utilization in several European countries. When the utilization is high, there is a high probability of delay propagation, which leads to a lower punctuality.

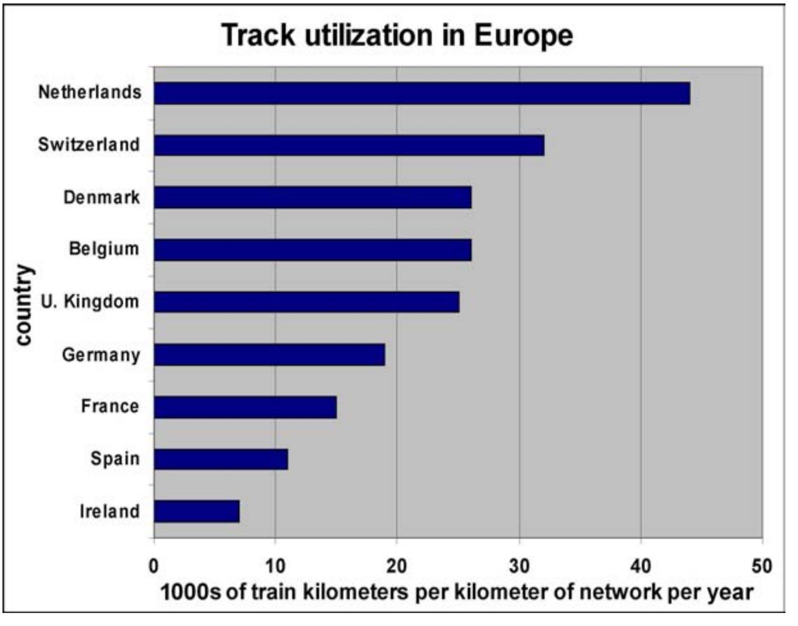


Figure 1.5: Track utilization in Europe (Poort, 2002).

Another utilization comparison could be made with dense metro networks. In fact the London Underground and the Metro of Paris have utilizations, which are 3 and 4 times higher than the Dutch network, respectively (151,000 and 208,000 train kilometers per network kilometer per year). However, metro systems have at least three advantages over national railway networks. First of all, minimal headways are much smaller in metro systems, due to the much smaller maximum speeds and resulting breaking distances. Secondly, metro lines are usually physically independent of each other, which simplifies operations. Furthermore, as opposed to heavy rail systems, most metros are operated with the same stopping pattern. This means that metros can run closely behind each other over longer distances. With the speed differences that exist in national railway traffic, this is impossible.

A utilization comparison of the Netherlands can be made with the German state North Rhine Westphalia. Table 1.1 shows that they have a comparable size and population. The Netherlands has a large population concentration in the western part of the country, just like North Rhine Westphalia has in the Ruhr area. The number of train kilometers in the Netherlands is 18% higher than in North Rhine Westphalia. However, the railway network of the Netherlands is only 53% of the network of North Rhine Westphalia. This means that the Dutch network is more than twice as crowded.

	Netherlands	N.R.Westphalia
land area (square kilometer)	34,000	34,000
population	16,000,000	18,000,000
large urban area	Randstad	Ruhr area
train kilometers per year	131,000,000	111,000,000
railway network length (km)	2,811	5,300

Table 1.1: Comparison of the Netherlands and North Rhine Westphalia.

Punctuality Improvement The foregoing should make clear that it is important for railway companies to achieve high punctuality figures. This is not only to fulfill their contracts, but also to attract passengers for sustainable operations in the future. Several options are open to improve punctuality.

- **External factors** Much of the unreliability of the railway services is caused by external factors, such as suicides, accidents and bad weather conditions. Unfortunately, these factors can hardly be influenced. However, changing level road-rail crossings into level free crossings is a way to reduce external causes of delay, but it is quite costly.

- **Infrastructure** To decrease the interdependencies between different trains, it can be helpful to build new infrastructure. Spare infrastructure can also be very helpful in case of disturbed train traffic. However, infrastructure extensions are very expensive. This is not about millions, but about hundreds of millions or even billions of euros.

  Besides the physical presence of the infrastructure, it also has to be dependable. In order to reduce the failure rates of the railway infrastructure, an intensified preventive maintenance process is required.

- **Rolling Stock** Dependable rolling stock may certainly improve the reliability of a railway system. However, new rolling stock requires large amounts of money, and is not necessarily more reliable. Increased preventive maintenance could lead to more dependable rolling stock.

- **Timetables** Constructing timetables with less dependencies between the train services can also increase the punctuality. However, this must not hurt other important timetable characteristics such as planned travel times. Moreover, it is not at all straightforward how to construct timetables with less dependencies.

This thesis aims at the last point: finding ideas and methods to construct timetables which are more reliable. The precise aim of this research is described in the next section.

A further discussion of punctuality, disturbances, track utilization and passenger perception can be found in Dekker and Vromans (2001) and Palm (2004), among others.

1.3 Aim and Scope of the Thesis

The aim of this thesis is to develop planning rules, which explicitly deal with punctuality. For most changes in railway planning it is quite clear whether they have a positive or a negative influence on punctuality. Some of them, however, have both a positive and a negative impact. Then it may be hard to tell which influence is the dominant one. Even more importantly, it is very hard, if not impossible, to quantitatively relate planning characteristics to punctuality and stability.

In addition, a number of planning rules at NS evolved over time to what they are today. Although they are based on 160 years of experience, they have never been verified scientifically.

In the whole search for a high punctuality, the passengers should not be forgotten. Some measures to improve punctuality have no or little effect for passengers besides a better reliability of the timetable. However, other punctuality increasing measures may also have passenger unfriendly effects. It can be imagined, for example, that spreading trains over the hour at a transfer station decreases the interdependence of the trains at that station. In that way, punctuality will probably rise, but transfer times will also increase.

> ***The central issue in this thesis is to develop rules and instruments for supporting the generation of more reliable timetables.***

This issue is so extensive and complicated, that it should be tackled piecewise. In a later stage small sub-problems can either be extended or combined.

The following five research steps will guide the research towards a better understanding of punctuality, reliability, stability and robustness.

1. **Describe determinants of punctuality** Describe the influence of planning rules and other aspects on punctuality.
2. **Estimate the impact of different aspects on punctuality** Quantify the influences of planning characteristics on delays and punctuality.
3. **Compare the quality of different timetables** Develop models or tools which can be used to compare timetable scenarios based on stability and robustness.
4. **Apply the theoretical models in practical cases** Use these models to verify the appropriateness of current planning rules applied by railway companies with respect to punctuality, stability and robustness.
5. **Improve the timetable** Develop generic models to construct and optimize timetables with respect to delays and other quantitative objectives.

For all practical implications of these steps, the Dutch situation will be studied, and, more specifically, that of NS, as a reference.

Furthermore, note that only "small" disturbances can be taken into account when planning railway traffic. It is impossible to develop a reasonable timetable which is capable of absorbing large disturbances. Additionally, dispatching rules, which are generally applied in case of large delays, are often unpredictable and hard to model. Therefore the research focuses on small delays only.

1.4 Relevance of the Thesis

1.4.1 Scientific Relevance

A railway system can be considered as a very large and complex stochastic dynamic system. This system is already interesting by itself. Theory on stochastic systems commonly concerns simple systems. Large, complex systems are usually analyzed with deterministic methods. Many authors have reported on deterministic optimization of railway systems over the last twenty years. Models including the stochastic behavior of railway traffic are only emerging now. The complicated relations and the uncertainties in railways lead to scientifically challenging problems in the field of transportation science.

One of the objectives of this research project is to give insight into the appropriateness of certain performance indicators for a priori evaluating the logistic plans underlying such a stochastic dynamic system. The appropriateness of these performance indicators will be judged on their capability to guide the design of a stochastic dynamic system with a high robustness and a high stability. Since general knowledge on such performance indicators hardly exists, it is a challenging scientific problem to create this knowledge.

Furthermore, a stochastic dynamic system can be modeled based on several alternative modeling approaches, such as simulation or queueing models. Each approach has certain advantages and disadvantages. This research project will give more insight into these advantages and disadvantages. In this thesis, an important criterion will be the practical applicability of the alternative approaches. The final achievement of this thesis is a symbiosis of these methods. The resulting model is a powerful timetable optimization tool.

The scientific relevance of our thesis is stressed by the scientific appreciation of the joint papers *Reliability and Heterogeneity of Railway Services* (Vromans et al., 2006), and *Stochastic Optimization of Railway Timetables* (Vromans and Kroon, 2004). These are based on the research described in Chapters 5 and 6, respectively.

1.4.2 Social Relevance

As mobility is increasing and congestion is a large problem in the Netherlands, railway traffic will be increasingly important in the near future. The immense media attention for NS over the last period mirrors the public interest in high quality public transport.

Any improvement in public transport is good for the society as a whole. It will not only be beneficial for the user, but, with an increasing number of passengers, the non-users will find less cars on the road. The accessibility of the urban areas in the Netherlands largely depends on the success of public transport. Besides that, the environment benefits from public transport as well. To attract new passengers, the reliability of trains and buses has to be high. This mainly concerns the punctuality and the upholding of connections. The latter one is a result of the former.

1.4.3 Managerial Relevance

The managerial benefit of this research lies in the possibility for railway management to take better supported decisions concerning punctuality. Punctuality seems to be one of the most important features of railway transport to attract more passengers. But more importantly, NS have to pay millions of euros in fines to the government when they do not reach a certain level of punctuality. Furthermore, an agreement with the passenger organizations allows for additional ticket price increases.

There are many possibilities for the improvement of punctuality. For example, after the disastrous fall of 2002, when 7% of all trains had to be canceled, a 17 million euro project was initiated to improve the punctuality. However, well-founded cost-benefit analyses are scarce, despite the enormous investments that are involved in some projects aimed at improving the punctuality. Decisions should be based on a quantitatively supported trade-off between the investments and the improvements in punctuality, other services and travel-characteristics.

Furthermore, our research has led to a better understanding of timetabling norms. Applying these insights will lead to more robust timetables, or give the current norms a scientific foundation.

1.5 Outline of the Thesis

Chapter 2 introduces the railway planning process to the reader, and more specifically the Dutch situation. It then also describes disturbances and delay propagation, and how planning can help to improve reliability. Furthermore, performance measures and perception are discussed. Finally a literature review is given.

In Chapter 3, existing timetabling models and timetable evaluation models are discussed. The first part focuses on timetabling models. More specifically, the timetabling tool DONS is discussed. Later, two types of timetable evaluation models are discussed. First, the focus is on models based on max-plus algebra. After that, simulation is addressed as a way to evaluate timetables. In particular, the simulation tool SIMONE is discussed.

In the following chapters several timetabling characteristics are investigated. Chapter 4 focuses on the distribution of running time supplements. This problem is researched by an analytical model, a numerical approach and simulation. All these methods show that the proportional supplement allocation used in practice is not optimal with regard to average arrival delays.

This is followed in Chapter 5 by a closer look at the influences of heterogeneity of railway traffic on the reliability. First, new heterogeneity measures are developed. Then both a theoretical and a practical simulation case are worked out. These cases both analyze the influence of the heterogeneity on reliability, and the appropriateness of the new heterogeneity measures.

In Chapter 6, an innovative stochastic optimization model is introduced. To the best of our knowledge, this is the first optimization model that takes the delay propagation explicitly into account. The model is able to decrease the average delay of existing timetables considerably within the model settings. The model can be used in different settings, of which the standard version can handle a large subnetwork.

Chapter 7 is the concluding chapter. First the results from the earlier chapters are discussed. Then we relate these results to the research questions, that are described in Section 1.3. Finally, some recommendations for further research are given.

Chapter 2

Timetabling and Delays

In a railway system, a timetable is a list of times when trains are supposed to arrive at or leave from a certain location. The timetable is a promise of the operator to the passengers, telling them when and where trains are planned to run.

The time difference between a departure at one station and the arrival at the next station is the running time between these two stations. The time between the arrival at a station and the consecutive departure from the same station is the dwell time. This dwell time enables passengers to alight or board the train. In general, public transport timetables are rounded to minutes.

This chapter starts with an introduction on timetabling. This is followed by a description of timetabling (Section 2.2) and dispatching practice (Section 2.3) in the Netherlands. Section 2.4 deals with disturbances. The interdependencies in railways and their resulting secondary delays are described in Section 2.5. Then Section 2.6 is dedicated to the implication of some important planning principles related to reliability. Furthermore several performance measures are described in Section 2.7. Finally, in Section 2.8 a literature overview on railway planning and timetable reliability is presented.

2.1 Introduction to Timetabling

Both for passengers and operators it is important to have timetables in public transport. The travelers use the timetable to plan their trips. For operators the timetable forms the basis for further planning, such as rolling stock planning and crew planning. For high frequency services, such as metros, timetables are sometimes not published externally, but they do exist internally at the operator for planning purposes. A possible sequence of planning processes at a railway operating company is given in Figure 2.1. This figure is followed by a brief description of each of these planning phases. The main focus of this thesis is on timetabling.

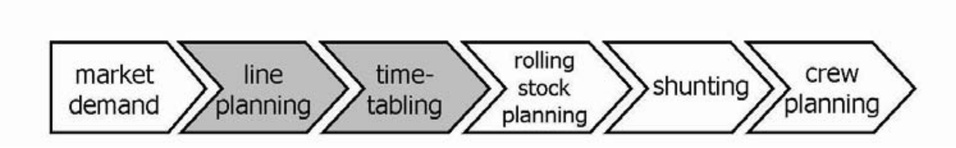


Figure 2.1: Sequence of interdependent railway planning phases.

Market demand and line planning In railway planning it is important that travel demand and travel possibilities match. Starting from an estimated market demand, usually represented by an Origin-Destination Matrix, a line plan is worked out. This line plan includes the train lines, but no departure and arrival times. A train line consists of a departure station, an arrival station and the intermediate dwell stations. In general, a line is also defined for the opposite direction. Depending on the travel demand, it is to be decided how often a line has to be operated.

Timetabling When the line plan has been worked out, timetabling can proceed. For all train lines, departure and arrival times are defined at all stations which are passed. When timetabling the train lines, many restrictions and preferences have to be taken into account. A selection is described in the following.

First of all, technically minimal running times and dwell times have to be respected, to be able to operate the timetable. The technically minimal running time from A to B is defined such that the used type and amount of rolling stock can run from A to B in at most this technically minimal running time in at least 95% of the circumstances while obeying the safety system. However, this is hard to determine. In practice, the running time under 'normal' circumstances is calculated from the technical characteristics of the train or from test runs. Circumstances that influence the running times include wind speed, deviations in overhead power supply, the number of passengers and the train driver's behavior. Some examples of these circumstances are described in the following intermezzo (Gielissen, 1983).

Influences on the running time Many circumstances influence the running time of a train. Although the maximal speed can almost always be reached by a train, several circumstances can have a sizable influence on the acceleration rate of a train. The examples of time losses and time gains mentioned here are based on an acceleration from 0 to 140 km/h without any speed limitations.

The regular *overhead power supply* in the Netherlands is 1500 V. However, a decrease of power from 1600 V to 1400 V increases the time for accelerating from 0 to 140 km/h with 15 seconds.

Accelerating from 0 to 140 km/h with a *headwind* of 40 km/h instead of no wind at all makes a difference of about 5 seconds.

The *weight* of a train has also a large influence on the acceleration: 105 metric tons instead of 85 metric tons (a difference of about 250 passengers) costs around 10 seconds for accelerating from 0 to 140 km/h.

The *accuracy of the top speed* is another point of discussion here. Running a train at 130 km/h instead of 140 km/h implies a time loss of 2 seconds per kilometer; running 120 km/h costs even 4.3 seconds per kilometer. At lower speeds, a 10 km/h deviation implies even larger time losses: 30 km/h instead of 40 km/h extends the running time by 30 seconds each kilometer.

Two final notes have to be made. First it has to be recognized that the acceleration loss is incurred every time that the train has to accelerate: for a local train with many stops these acceleration losses can easily add up to several minutes. Secondly, note that a combination of negative circumstances may lead to more time loss than the sum of individual time losses may indicate.

Planned running times, which are rounded to minutes, between large nodes in the network have to be larger than the technically minimal running times. The positive difference between the planned running time and the technically minimal running time is called *running time supplement*. This difference does not only emerge from rounding, but is also put into the timetable on purpose to absorb possible small delays. Over short distances there are sometimes negative running time supplements. These have to be compensated by positive supplements before a large node is reached. The allocation of running time supplements in timetables is discussed thoroughly in Chapters 4 and 6.

Many infrastructural or safety constraints have also to be respected in timetabling: a track, a switch or a platform can only be reserved and used by one train at a time.

Because of safety, it is necessary to have a certain distance (or headway) between the trains, which is determined by the safety system.

Furthermore, there are many market demands. These are abstracted from the passenger demand. Due to the fact that the line plan does not offer direct connections for all passengers, passengers sometimes have to transfer. Besides the time that is needed to go from one train to another, these transfer times should be scheduled as short as possible, to avoid excessive waiting times.

An additional passenger preference is explained in Figure 2.2. Assume both lines, A-B-C and A-B-D, are operated once per hour. This implies that there are two services per hour from A to B. It is desirable to have these services in 30-minute intervals, and not at, for example, alternating headways of 15 and 45 minutes.

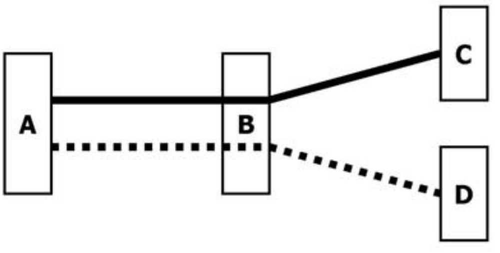


Figure 2.2: Two lines, sharing the same trajectory from A to B.

Finally, there are logistic preferences which originate from the operator. For example, to have a more efficient timetable, layover times at line endpoints should not be too long. When the train crew has to stay with the same rolling stock as much as possible, that would not only imply long idle times for the rolling stock, but also for the crew.

In practice, timetabling is made up of two different problems: networkwide timetabling of the tracks and timetabling the platform occupation per station. For timetabling time-distance diagrams, such as in Figure 2.3, are used, which depict the trains over a certain track or set of parallel tracks. Distance is represented horizontally, time vertically, starting from the bottom. The solid vertical lines are stations, the dashed vertical lines are other timetable points. The trains are represented by the diagonal lines: the flatter they are, the faster the train. Vertical jumps in the train paths are dwell times.

For the platform assignment in stations, platform occupation charts, or POCs, are used. A POC only includes one station. An example for Gouda is given in Figure 2.4. Each horizontal line in this graph represents one platform, from minute 0 to 60 for each cycle. All trains occupying one of the platforms in Gouda are depicted

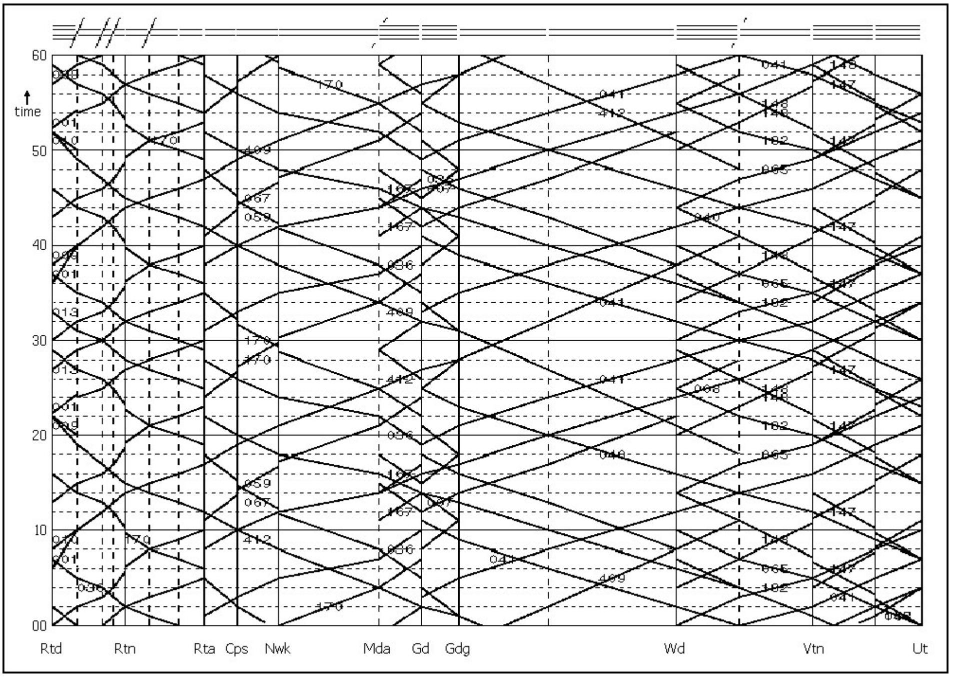


Figure 2.3: Example of a time-distance diagram for Rotterdam – Utrecht.

at the corresponding platform. For each train the origin, the destination and train line number are given. Sometimes a platform is separated in an A- and B-part (here platforms 2, 3 and 5). Trains occupying the A-part of the platform are depicted just above the horizontal platform line, the trains occupying the B-part are drawn just below this line. When a train occupies both parts of the platform, see for example line "020" at platform 5, a double line is used. The trains at platform 11 come from a shunting track.

As argued above, rolling stock and crew planning are kept in mind during the timetabling phase. However, the real rolling stock schedules are based on the finished timetable. Not all consequences of the timetable on the rolling stock circulation can be foreseen in the timetabling phase.

Rolling stock planning For each train in the timetable, rolling stock is needed. In this phase both the amount of rolling stock and the type of rolling stock are important. The amount depends on the expected number of travelers, the type depends on the train service. For example, a local train which has to stop every few minutes needs rolling stock with high acceleration and deceleration rates. For long distance trains with larger travel times more comfortable train units are preferred.

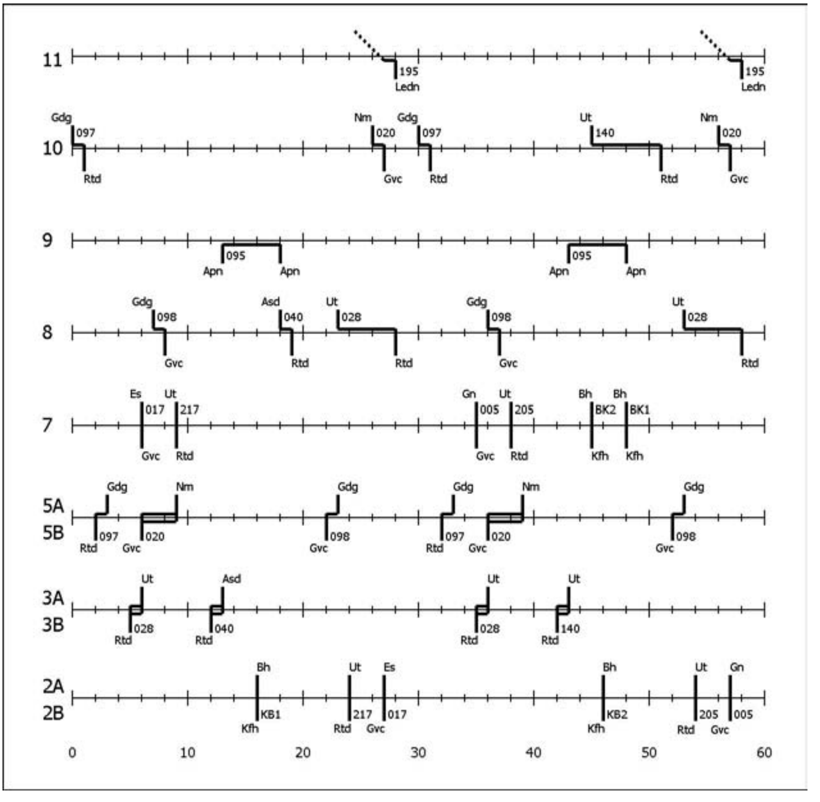


Figure 2.4: Example of a platform occupation chart for Gouda.

Shunting Not all train units are needed all day long. More units are needed during rush hours, and hardly any are used during the night. Therefore, trains are shortened at the end of the rush hours and extended at the start of the rush hours. To execute these shunting movements from platforms to shunting yards and the other way around shunting plans are needed. At the end of the services almost all units are shunted to the shunting yard. At the shunting yard, also all kinds of movements are needed for cleaning and maintenance of the trains, possibly disturbing the system.

Crew planning When all the train and shunting movements have been scheduled, the crew planning can start. Each train needs at least one driver. Depending on the length of the (physical) train and other characteristics, a certain number of conductors has to be assigned to each trip. Additional crew is needed for shunting movements.

2.1.1 Cyclicity

In the Netherlands and in many other countries, a cyclic timetable is used. This means that all train lines are operated with some fixed interval time, the cycle time. In the Netherlands a cycle time of one hour is used, leading to a timetable based on a One-Hour Timetable or OHT. An OHT is a timetable for one hour, which is constructed in such manner that the trains in minute 60 run exactly the same as in minute 0, thus being able to run the trains precisely the same as one hour before.

The advantage of a cyclic timetable is probably largest for the passengers: they only have to remember at which minute past the hour their train leaves and they know the connection for the whole day. For the operator the main advantage is that only one cycle has to be timetabled. The rolling stock schedules are in principle also cyclic, when the timetable is cyclic. Crew schedules are usually not cyclic.

Cyclicity of the timetable also has some disadvantages. First of all it is very inflexible: demand deviations over the day ask for different train intervals, which are fixed in a cyclic timetable. Secondly, an “easy” cyclic timetable has a cycle time of 30, 60 or maybe 120 minutes. When market demand asks for a deviating cycle time, this leads either to “difficult” cycle times or to large inefficiencies for the operator. Furthermore, the cycle time has to be equal, or an integer multiple, for all lines, to avoid complicating transitions between areas with different cycle times. This is true both for the planning of the operator and for the transfers of passengers.

The timetable for a whole weekday can by created by copying the OHT as many times as desired. This is shown in Figure 2.5. An example of a cyclic timetable for the so-called 800-line from Haarlem to Maastricht on a weekday is given in Table 2.1.

Note that NS use three OHTs already: one general OHT, one for the morning rush hours, and one for the afternoon rush hours. On top of that many deviations from the cyclic pattern are scheduled in practice. First of all, the rush hour OHTs are not applied for the same time intervals at all trajectories. The earliest and latest trains of a day often deviate from the OHT. The transitions from and to the rush hour OHTs need adjustments, too. Furthermore, there are often fewer trains during the quiet hours, such as the late evening or Sunday morning. Additionally in the Netherlands there are a few international train lines, which are operated only several times a day. Many other deviations from the OHT are present in the timetable.

2.1.2 Symmetry

Cyclic timetables are usually characterized by symmetry. For passengers, a symmetric timetable means that they have more or less the same traveling time from A to B as from B to A, they also have to transfer from one train to another at the same

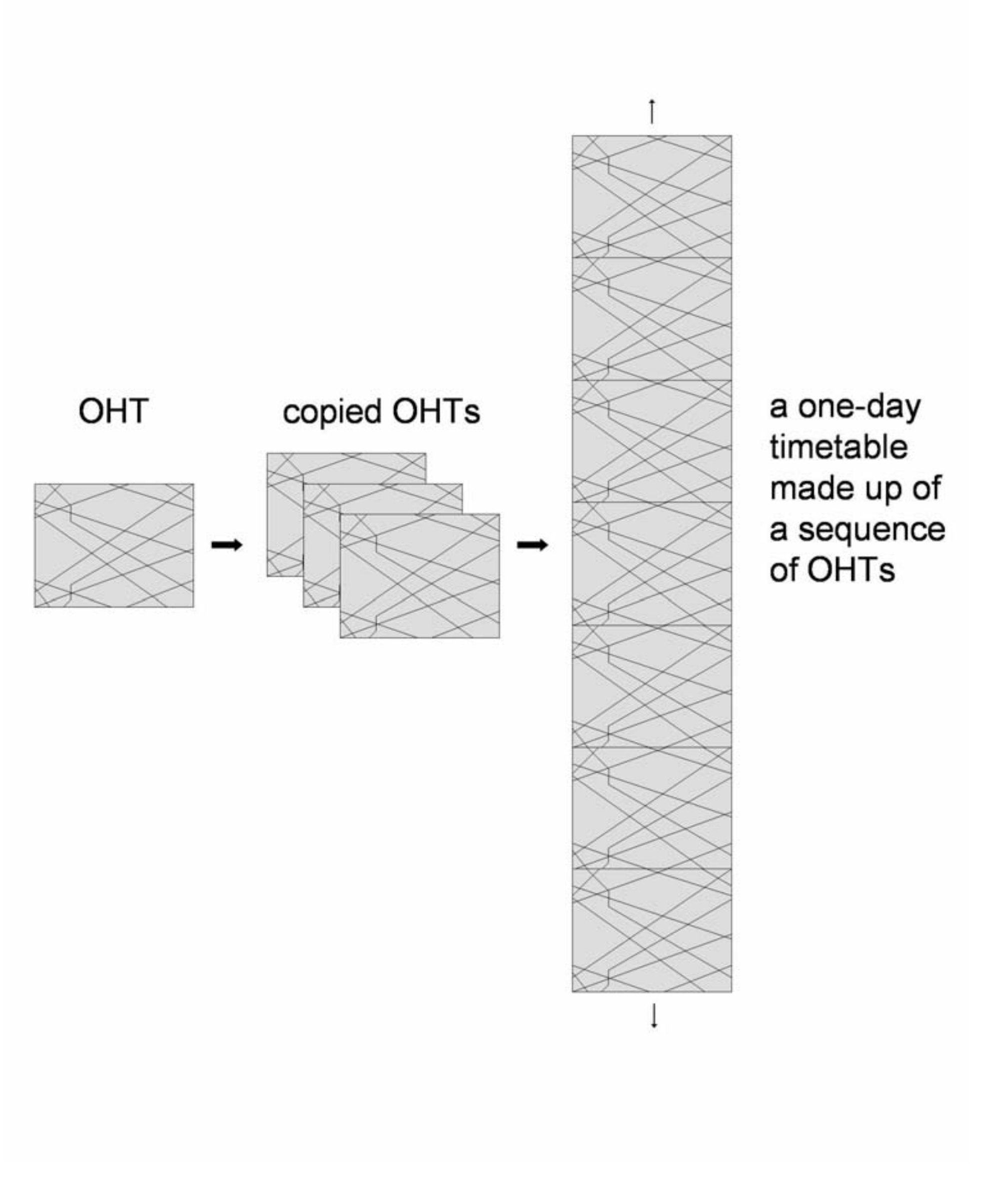


Figure 2.5: From a One-Hour Timetable to a timetable for one day.

train number		817	821	825	829		881	885	889
Haarlem	d		6.11	7.11	8.11	...	21.11	22.11	23.11
Amsterdam Centraal	a		6.25	7.25	8.25	...	21.25	22.25	23.25
Amsterdam Centraal	d		6.28	7.28	8.28	...	21.28	22.28	23.28
Amsterdam Amstel	a		6.36	7.36	8.36	...	21.36	22.36	23.36
Amsterdam Amstel	d		6.36	7.36	8.36	...	21.36	22.36	23.36
Duivendrecht	a		6.40	7.40	8.40	...	21.40	22.40	23.40
Duivendrecht	d		6.41	7.41	8.41	...	21.41	22.41	23.41
Utrecht Centraal	a		7.01	8.01	9.01	...	22.01	23.01	0.01
Utrecht Centraal	d		7.04	8.04	9.04	...	22.04	23.04	
’s-Hertogenbosch	a		7.32	8.32	9.32	...	22.32	23.32	
’s-Hertogenbosch	d		7.34	8.34	9.34	...	22.34	23.34	
Eindhoven	a		7.54	8.54	9.54	...	22.54	23.54	
Eindhoven	d	6.59	7.59	8.59	9.59	...	22.59	23.59	
Weert	a	7.17	8.17	9.17	10.17	...	23.17	0.17	
Weert	d	7.17	8.17	9.17	10.17	...	23.17	0.17	
Roermond	a	7.31	8.31	9.31	10.31	...	23.31	0.31	
Roermond	d	7.32	8.32	9.32	10.32	...	23.32	0.32	
Sittard	a	7.47	8.47	9.47	10.47	...	23.47	0.47	
Sittard	d	7.48	8.48	9.48	10.48	...	23.48	0.48	
Maastricht	a	8.03	9.03	10.03	11.03	...	0.03	1.03	

Table 2.1: Cyclic timetable for the 800-line from Haarlem to Maastricht, with d = departure time, and a = arrival time.

stations and the transfer times are more or less equal.

When train lines are defined in both directions, the running times and dwell times are equal in both directions, and all connections are defined in two directions, a timetable will automatically be *symmetric*.

Symmetry also implies that there are two *symmetry times* within one cycle for which a train line meets its counterpart in the opposite direction. This can be both at a station or somewhere at an open track. The symmetry time is equal for all lines. The two symmetry times are half a cycle apart. In the Netherlands, the symmetry times are approximately 16 and 46 minutes past the hour. When train lines have a frequency higher than one train per cycle, these lines will meet their opposite counterparts more often. In fact, a train line meets itself twice its frequency per cycle, if the travel time of this train line from begin to end is at least as large as the cycle time. With the One-Hour Timetable in the Netherlands, and many lines operating twice per hour, a train line encounters itself four times per hour, at approximately 1, 16, 31 and 46 minutes past the hour.

These symmetry times are often used to plan passenger transfers at large transfer stations: all trains from all directions arrive just before the symmetry time and depart just thereafter. This enables transferring from all arriving trains to all departing trains. A station with this kind of transfer possibilities is referred to as a symmetry node. Good examples of symmetry nodes in the Netherlands are Utrecht Central (around 1, 16, 31 and 46) and Zwolle (around 16 and 46). Table 2.2 shows the symmetric arrival and departure times for Zwolle.

			16-symmetry				46-symmetry			
train line			from	a	d	to	from	a	d	to
intercity	500	north	Gvc	15	17	Gn				
splitting			”	”	19	Lw				
intercity	500	south	Lw	11	18	Gvc				
combining			Gn	14	”	”				
intercity	700	north					Shl	46	49	Gn
splitting							”	”	52	Lw
intercity	700	south					Lw	40	48	Shl
combining							Gn	44	”	”
interregional	3600	turn	Rsd	09	21	Rsd	Rsd	43	53	Rsd
interregional	3800	turn	(8000)		33	Emn	Emn	37		(8000)
interregional	5600	turn	Ut	09	21	Ut	Ut	40	51	Ut
local	7900	turn	Es	12	21	Es	Es	42	51	Es
local	8000	turn	Emn	08		(3800)	(3800)		03	Emn
local	8500	turn	Kpn	15	20	Kpn	Kpn	44	51	Kpn
local	9100	turn	Gn	03	25	Gn				

Table 2.2: Symmetry in Zwolle: around 16 and 46. d = departure time, and a = arrival time. The intercity lines have Zwolle as an intermediate stop, going north or south. The other lines have Zwolle as their terminus and turn around onto the same line in opposite direction; only the 3800 and 8000 (from and to Emmen) turn on each other. For the station abbreviations see Appendix A.3.

Traveling times between symmetry nodes have to be just below a multiple of half the frequency interval. The time remaining up to the multiple of half the frequency interval is the time which is available for transfers. For example, both the intercity trains and the interregional trains between Utrecht and Zwolle run every thirty minutes, implying the running time difference to be a multiple of fifteen minutes. Indeed, the intercity trains have a running time (including the intermediate dwell times) of 53 minutes. They leave Utrecht just after 16 and 46, and arrive in Zwolle just before 16 and 46 (one cycle time later). The interregional trains have a running time of 68 minutes; they leave Utrecht just after 01 and 31, and arrive in Zwolle just before 16 and 46. In the opposite direction, all trains leave just after 16 and 46 from Zwolle, such that the intercity trains arrive in Utrecht just before 16 and 46, and the interregional trains just before 31 and 1.

Small running time and dwell time differences cause the symmetry time to deviate a little bit from location to location. Some lines with only one train per hour have the opposite direction planned in the 'wrong' half-hour, which causes the symmetry times to be 1 and 31. As long as all connecting services either also have a symmetry of 1 and 31 or have a frequency of two trains per hour, this does not cause any practical problems for passengers. The choice for the 'wrong' half hour mainly originates from the international trains and the symmetry around 0 in Belgium and Germany. The last minute is forced into the timetable by adjusting the running time or dwell time supplements around the border.

2.2 Timetabling at Netherlands Railways

In this section the logistic planning process at NS is described. The basic ideas behind the planning process at other railway operators are similar, but differences in the implementation may occur.

In the Netherlands, railway infrastructure ownership is separated from the railway operators. This means that the operators do not own the tracks, and the same infrastructure is often used by multiple operators, such as the passenger operators NS, Syntus and NoordNed, and the cargo operators Railion, ACTS and Rail4Chem. Furthermore, capacity is reserved for infrastructure maintenance.

The infrastructure is managed and maintained by ProRail. NS are by far the largest operator in the Netherlands. In dialogue with other operators, both for passengers and cargo, a timetable is worked out by NS. ProRail has to settle the last disputes between the different operators, but more importantly, it has to check the feasibility and safety of the proposed timetable.

In the following, only the planning process at NS is described. This planning process starts very roughly, to be worked out in more detail in later stages, as Figure 2.6 shows. These details are added both on the infrastructural level, from national lines to local shunting movements, and in time, from a single One-Hour Timetable to an elaborated plan for each specific hour of each specific day.

2.2.1 Central and Local Planning

Timetable planning at NS is separated between the central planning department in Utrecht and the local planning departments at several locations throughout the country. Centrally, a national timetable is worked out. This step in the planning takes both the capacities of the trajectories and the stations into account. However,

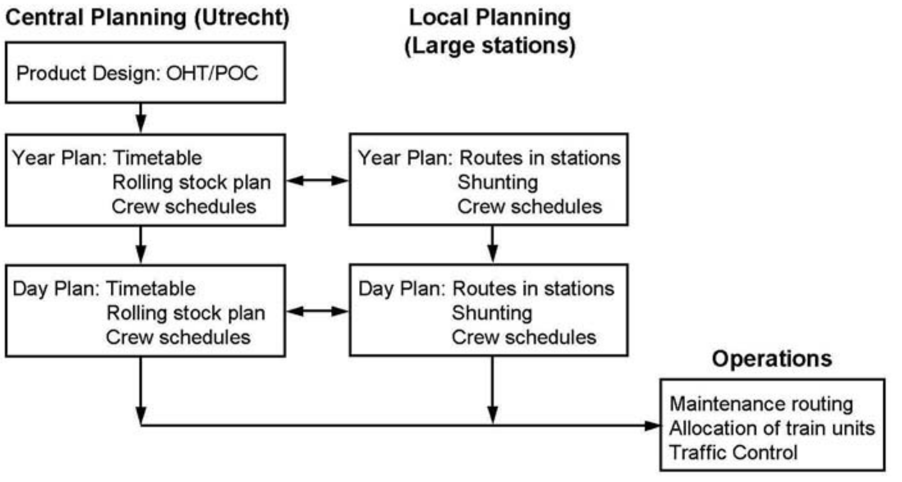


Figure 2.6: Timetable planning process at NS.

the details of the platform occupation and the routing through the stations have to be worked out locally. Shunting movements have to be scheduled also. When a local office is not able to schedule the proposed timetable, iteration loops with the central planning department take place. Usually only small adjustments are needed in these steps.

2.2.2 From OHT to Daily Plan

The central planning department starts with the development of a One-Hour Timetable (OHT). In practice this OHT is largely based on the OHT of the year before. Still, there will always be changes in the OHT, inspired by shifts in market demand and altered infrastructure. Parallel to the OHT, planners also work on platform occupation charts (POCs) for the appropriate stations (see Section 2.1).

Besides a general OHT, an extended OHT is developed for the rush hours and other OHTs are created for the evenings and weekends. Besides the timetable, these OHTs include information about the standard hourly rolling stock circulations. This may also include the regular splitting or combining process of train-units at a certain station for a certain line.

The different OHTs are then evaluated by the local planners. They mainly focus on the large stations and the shunting movements there. Routing the trains to and from the platforms is sometimes not possible given the OHTs and the accompanying hourly rolling stock circulation. These problems have to be solved in an iterative process with the central planning department.

When the OHTs have been approved by the local planning departments and by ProRail, they are worked out to a 7×24 Plan or Weekly Plan. There have not been rigorous changes to the core of the timetable over the last decade. Therefore, the old Weekly Plan is used as a basis, and transformed into the new Weekly Plan.

Especially the transitions from and to the rush hours require non-hourly shunting movements, or in other words, train movements which were not included in the OHTs. Because these additional movements often have a large impact at the stations, local planning is important in checking the possibilities in this phase.

In the Netherlands, the timetable is usually valid for one year, which means that the Weekly Plan is, apart from certain holiday periods, repeated one year long. However, for every single day of the year adjustments are necessary. The Daily Plan for example includes extra trains for events, altered schedules during infrastructure maintenance, adjustments for national holidays, and reduced traffic during vacation periods. In the end, usually not one pair of identical days or even identical hours can be found anymore. Again, first a rough Daily Plan is created by the central planning department, which is later worked out in detail by the local planning departments with possible iterations between the two.

The Daily Plan is communicated to the Rail Traffic Control at least 36 hours before the actual operations.

2.3 Operations and Dispatching

During the daily operation of trains, stochastic disturbances influence the railway traffic. These disturbances lead to smaller and larger delays.

Dispatching organizations monitor the stochastic railway traffic. In the Netherlands both the infrastructure manager ProRail and the operators have their own dispatching organization.

As long as delays are small, dispatchers will not intervene. Only in some cases two connecting trains will wait to enable passengers to transfer.

However, larger delays and partial or full blocking of tracks do need intervention of dispatchers. The dispatchers will attempt to minimize the consequences of the delays and blockades for the railway traffic and the passengers.

The four *Traffic Control Centers* (or TCCs) of the infrastructure manager ProRail are responsible for the redistribution of the remaining infrastructure between the operators. For many situations there are standard dispatching strategies. For other situations, solutions have to be created on the spot. The final decisions are made in consultation with the operators. Initiating and communicating the decisions is the main task of the TCC.

NS has its own four *Dispatching Centers*, corresponding to the TCCs. They have to analyze the changes in the timetable and make adjustments to the crew and rolling stock schedules. The Information Distribution Center, part of the Dispatching Center, informs train drivers and conductors about the changes.

ProRail also has thirteen *Train Regulating Offices*, where the decisions taken by the Traffic Control Centers are carried out. They regulate the train traffic by operating the signals and switches. NS has its own thirteen *Coordination Organizations* to coordinate the changes in train traffic. This mainly includes the arrangement of additional shunting movements and the coordination of the adjusted departures. Service employees of NS are present at the stations to inform passengers.

The separation of responsibilities implies that a good cooperation between the organizations is necessary. Considering the time pressure under which dispatching decisions have to be taken, it is even more obvious that smooth communication between the different parties is important. Olde Hartman (2003) describes the roles of different dispatching organizations in more detail.

This thesis aims at reducing the propagation of delays by using smart planning rules. However, no reasonable railway plan is robust or stable enough in case of large delays or the blocking of tracks. Therefore we only regard small delays in the remainder of this thesis. Traffic control organizations and dispatching strategies do not play a key role in this situation.

2.4 Disturbances and their Causes

Disturbances are mistakes, malfunctions or deviating conditions within the railway system or its environment, which influence the railway traffic. Disturbances have many different causes, for which many different railway organizations are responsible. Moreover, a sizable portion of the disturbances has an external cause.

The delays caused by disturbances are called primary delays. Secondary delays, or knock-on delays, are delays which are caused by delays of earlier trains. Due to the interdependencies in railway systems, which are described in Section 2.5, a large part of the delays are secondary delays.

2.4.1 Size and Distribution of Disturbances

Obviously not all the disturbances of railway traffic are equal in size. Some just cause a few seconds of delay, whereas others cause a trajectory or node to be blocked for several hours or even more.

A very common small disturbance is where a train departs just a few seconds late from a station. Although most people will not even realize the late departure, those seconds are taken from the scheduled running time, and the probability of on-time arrival will decrease. An energy black-out can cause massive delays and the cancelation of many trains during an extensive period of time. Anything in between is possible as well.

It is very complicated to determine the distribution of disturbances. All train detection data includes both primary and secondary delays. Analysis of total delays is already complicated, but it is even more difficult to separate primary and secondary delays, and to extract the disturbances.

Goverde et al. (2001) performed a detailed analysis of realization data in the Netherlands. They found that late arrivals, departures and dwell time prolongations all fit well to exponential distributions. However, these results are based on a limited period of realization data and on one location, Eindhoven, only. Yuan (2001) performed a similar analysis for another station, knowingly The Hague HS.

Due to the lack of further knowledge about disturbances, we mostly use exponentially distributed disturbances in this thesis.

2.4.2 Sources of Disturbances

Sources of disturbances in railway traffic are plenty. A selection of these sources is given here.

- **Planning** Planning may lead to some disturbances. A first example is just after the release of the new yearly timetable. Small mistakes may have slipped into the plans, because most present day planning systems have hardly any conflict signalling. Another example is where the used characteristics of new types of rolling stock (which are supplied by the manufacturer) overestimate the real possibilities. Another source of disturbances is where longer and heavier trains are planned and operated during rush hours than those on which the running times were based. Also dwell times may be too short for the rush hours. In general, planning only leads to marginal disturbances and primary delays.

- **Infrastructure** Failures of the infrastructure are a reason for large delays. Infrastructural disturbances include malfunctioning switches, broken catenary, failing signals and non-working automatic barriers. These failures have their effect on all trains which are planned to pass the infrastructure in question. A separate reason for disturbances and canceled trains are maintenance works which take longer than planned. The power supply, not only for the catenary, but also for the Traffic Control Centers, is in hands of energy companies. Power black-outs are not common, but can have an enormous impact on the railway traffic over large areas and a long time span.

- **Rolling stock** When rolling stock breaks down it can block a trajectory or part of a station. Still, many small rolling stock malfunctions only have a minor influence on the railway traffic. Rolling stock problems include engine break-downs, leaks in the hydraulic system, problems with closing the doors, problems with splitting and combining train sets, and so on.

- **Human factors** Operating the trains is still mainly a human process and therefore it can not be flawless. For example, driving a train is a process with a stochastic nature. Each train driver has its own driving behavior, and even the same driver does not realize the same running time for the same track every time. Although one can think of many possible human factors, the frequency and extent of human errors is usually small.

- **Accidents with other traffic** Accidents happen quite regularly at level rail-road crossings, both at guarded and unguarded crossings. Furthermore approximately 200 suicides are committed on railway tracks every year in the Netherlands. Besides their severe direct influence on railway traffic, both accidents and suicides have a large mental impact on train drivers.

- **Vandalism** Also vandalism has to be included in this list. Regularly stones, trash or bikes are found on the tracks. Too often these items are not found until a collision is unavoidable, damaging both the tracks and the rolling stock. Additionally, coins on rail joints can cause signaling problems.

- **Passengers** Although passengers should not be blamed for delays, they can sometimes be seen as a disturbing factor. Passengers arriving on the platform just at the moment that the crew wants to close the doors often causes the conductor to wait a little longer. Also dwell times are sometimes extended because of pushing and pulling at the train doors, or even blocking the people trying to get off the train. Also *aggressive passengers* can cause delays. Sometimes it is even necessary for the conductor to call for police assistance. Furthermore, emergency breaks are sometimes pulled for fun.

- **Weather conditions** Weather conditions are another source of problems. An extremely fast temperature increase can cause the track to bend or cause signaling failures. An extremely fast decrease in temperature can cause the tracks to crack. The most problematic weather conditions usually occur during the fall, when a combination of leafs and moist causes the tracks to be very slippery. This causes breaking distances to increase and acceleration rates to go down. Although the resulting delays are usually small, under bad conditions they occur frequently and over large areas.

- **Along the track** Furthermore, playing children, strolling people or escaped cattle along the track force the train traffic to slow down to 30 or 40 km/h.

2.5 Interdependencies and Secondary Delays

Many disturbing factors and causes for primary delays were mentioned in Section 2.4, and these have a direct impact on the train services. However, the total number of delays is much larger due to delay propagation. Delay propagation is the spread of delays in the railway system, both in time and space. Delays spread around due to dependencies between train services. When a train is delayed because of another delayed train, this is a secondary delay. The main reasons for delay propagation are discussed in this section.

- **The train itself** If a train incurs a delay early on, and it is not possible to recover this delay, it will have this delay up to its endpoint. In case of long distance trains, delays may be carried along several measuring points. This means that one primary delay can easily be propagated to the other side of the network. However, we are still talking about one train: there is no 'real' secondary delay, but the primary delay is measured more than once.

- **Infrastructural use** The capacity of the railway infrastructure is limited. The Dutch railway network is the most heavily used, with on average almost 50,000 train kilometers per network kilometer per year (Poort, 2002). The Netherlands also have the smallest network length per person in Europe.

  The high utilization of the infrastructure implies a short headway between the trains. Still, because of safety, there is a minimal time headway between trains. Depending on the situation and location, the minimal planning distance ranges from several seconds to ten minutes. Especially at and around large stations and on the busy tracks in the western part of the Netherlands, planned headways between trains are often not much larger than the minimal headways. This is true for the open tracks, platforms at stations and level crossings.

The positive difference between the planned headway and the minimal headway is called buffer. When two trains are planned at a minimal headway, with no buffer time between them, and the first one is just slightly delayed, it will already cause the second train to be late: it incurs a secondary delay. When buffer is included in the planned headway, small delays do not directly lead to secondary delays.

In many cases a train is scheduled very close to many other trains. This means that a small delay can already cause many other trains to be late. These other trains can then cause further propagation of delays through the network.

- **Rolling stock circulation** When a train reaches its terminal station, its rolling stock will be used for a subsequent train. When this layover time is shorter than the delay of the arrival train, the second train will be delayed as well. When the rolling stock of one arriving train is used for more than one departing train, delay propagation may even go faster. When delays are large, and spare rolling stock is available, the rolling stock dependency between consecutive trains can be broken.

- **Crew schedules** Both train drivers and conductors change trains several times in their duties. When a crew member arrives late with a train early-on in his duty, he may transfer this delay to later trains. Note that both the driver and conductor must be in time to have a punctual departure.

  When additional spare crew members are available, they can be used when other personnel is late because of delays. Crew schedule dependencies can be reduced in this way.

- **Traffic control and dispatching** In case of small disturbances, traffic control does not have much work. Trains may run a little late, but nothing else has to be adjusted. Perhaps some local trains have to be retained to give priority to intercity trains. However, when larger disturbances occur, trains may have to be rerouted. Sometimes trains even have to be canceled partially.

  For a range of disturbed situations there are standard dispatching strategies. However, in many situations, dispatchers have to improvise. The implications of certain decisions are too complex to oversee in the short time span that dispatchers have. Especially canceling trains leads to additional shunting movements and adjustments in rolling stock and crew schedules. It is almost impossible to dispatch the railway traffic in an optimal way, given the delays which are already present.

- **Passenger connections** Dense railway networks, like in the Netherlands, are characterized by many passenger connections. In the Netherlands, these connections are often not more than 2 to 5 minutes. Within this small time window, passengers have to transfer from the feeder train to the connecting train. When such a connection is missed, passengers often have to wait thirty minutes. Connecting trains will therefore sometimes wait a little while when a feeder train is late. In the Netherlands, these waiting times are established in the WRT, the *waiting time passenger trains* (NS, 2004d). This states exactly for each connection, how long the connecting train should wait when the feeder train is late.

2.6 Planning and Reliability

Many timetable characteristics have a direct or indirect influence on the reliability. Furthermore, part of the interdependencies in the railway system are determined by the rolling stock and crew schedules. The ideas in this section are not supported scientifically, but most people in railway practice agree on the notions explained below. However, most of them are not documented, let alone quantified.

2.6.1 Line Plan and Timetable

The line plan and timetable form the basis of the railway planning. This is also the part of the railway plan that has to be communicated to the outside world (the customers), and between different companies in the railway sector.

The important project “Utilize and Build” (NS et al., 2003) within the Dutch railway sector is already mentioned in Section 1.1. This is the sector’s vision and the intended direction for the future up to 2020. To facilitate the ever growing demand for railway traffic, several timetabling ideas are studied in this project. Most of the ideas below are part of this project. However, many of the ideas of Utilize and Build to improve the reliability remain without a specific foundation.

Homogenization Speed differences play an important role in railway services. The current combination of local and intercity trains ensures that many stations are served (local trains), and long distances can be traveled in a relatively short time (intercity trains). This leads to *short travel times* for the large group of long distance passengers. However, as will be discussed in Chapter 5 and Section 6.8, speed differences often lead to an increased *delay propagation*. Homogenization of train traffic increases the *capacity of the system*. Furthermore, there are probably less peaks in the *capacity utilization of the platforms* at large stations. More discussion on the implications of homogenization can be found in Section 5.5.

Overtaking The *capacity of the system* can be increased considerably when slow local trains are overtaken more often by fast intercity trains. However, the local train experiences a certain *time loss* in that case, because it has to wait while it is overtaken. Repetitive overtaking leads to excessive travel times for local trains. Furthermore, this construction leads to a possible *delay propagation* in both directions: the local train disturbs the intercity train when it is late, but also the other way around. A large financial advantage of overtaking is that only small infrastructure investments are needed. Instead of going from two to four tracks for a certain line, a reasonable capacity increase can be attained by a few additional tracks at the overtaking stations.

Fixed corridors Due to the complicated network of train lines, more than 75% of the railway passengers can reach their destination without transfer in the Netherlands. On the other hand, this implies that when all trains at a certain part of the infrastructure are disturbed, they cause delays all over the network, because train lines that pass the disturbing part of the infrastructure run towards many parts of the network. In a simpler network with independent corridors, such as in metro systems, delay propagation between lines is almost non-existing.

Splitting and combining of trains A subject directly related to the fixed corridors is splitting and combining. The idea behind splitting and combining trains from and to different locations is that more passengers have a *direct connection*. Two trains partially traveling the same route can also be combined to decrease the use of the infrastructure. Instead of two smaller trains, both occupying their own time slot, there is only one longer train. This has a positive influence on the reliability. However, trains cannot always be split or combined within the scheduled time. Moreover, sometimes they cannot be split or combined at all due to technical problems: splitting and combining can have a *disturbing effect* on the operations.

Line lengths The length of a train line also has different influences on the qualities of a timetable. First of all, long lines can carry delays over long distances. But on the other hand long lines can provide many direct connections.

Supplements The analysis of supplements should be separated into two parts. First, it is important to know how much supplement has to be included in the timetable. In short, large supplements lead to longer travel times, but an increased capability to absorb incurred delays. Furthermore note that, given the Dutch safety system, the occupation of the tracks increases when the running times increase. The second question is where, and to which trains these supplements have to be allocated. A more extended discussion on the allocation of supplements can be found in Chapter 4.

Transfer synchronization Whether the timetable is based on fixed corridors or not, there will always be passengers that have to change trains to reach their destination. To have short travel times for transferring passengers, these transfers are often coordinated. A good connective network often has more than two trains connecting to each other at the same time. For example, in Zwolle six train lines have short connections to each other every 30 minutes (see Table 2.2). Although the station is used to its capacity at these symmetry times, the station is empty for the remainder of the hour: the overall utilization is quite poor. Still, the possible interaction between all trains around the symmetry times can lead to sizable delay propagation. Delays can also be propagated when trains wait for each other to uphold passenger connections. Note that the need for coordination of two connecting train lines decreases when the frequencies of the lines increase.

2.6.2 Rolling Stock Schedules

Besides the timetable, more planning processes influence the robustness of the complete plan. Some of the principles in rolling stock scheduling are discussed here. In fact, several characteristics of the rolling stock plan follow directly from the timetable. This is because the rolling stock schedule is based on the cyclic timetables, and is usually cyclic, too.

Rolling stock layover time The rolling stock circulation tells for each train trip with which composition of train units it has to be operated. The time between two consecutive train services for the same composition is the layover time.

Sizable delays can be absorbed with long layover times at line endpoints. However, this implies that expensive resources are idle for a long time. Furthermore, long layover times require much (platform) capacity at often crowded stations.

Layover locations Most non-ending train lines occupy a platform for no more than a few minutes. However, at line endpoints, trains have a layover time which is usually longer than a few minutes. At present, most train lines turn around at large stations. However, most large stations are down town large cities, where infrastructure is even scarcer than at other places. It may be advantageous to look for suitable locations for the line endpoints outside large cities. However, this should not lead to many empty train trips, or line extensions for which there is no or little demand. In the end this may lead to merging two short lines into one long line. In fact, turning around often also includes possible conflicts with other train lines, because the train has to cross several tracks before it can return in the opposite direction. This can be another reason to merge two train lines.

Spare rolling stock Spare rolling stock can be very useful in case of larger delays or the cancelation of train services. In such situations, the spare rolling stock can be used to operate train services for which the originally planned rolling stock is not available (on time). Large delays can be absorbed in this way. Although the availability of extra rolling stock can be very effective in restricting delay propagation, it is also costly. First of all, the rolling stock is expensive. Secondly, this spare rolling stock has to be available at the right place at the right time. Furthermore, not all types of rolling stock are suited to be used for all lines. Altogether, a good balance has to be found between the effectiveness of the spare rolling stock and its costs.

Rolling stock types The type of rolling stock that should be used for a train line depends on many characteristics such as service type, line length, expected number of passengers, and so on. Due to varying numbers of passengers over the day, or just the number of available train units of a certain type, two or even three types of rolling stock may be used on one train line. In the first place this may be confusing for the passengers, who often expect a certain type of rolling stock. Furthermore, in case of larger disturbances in the train traffic or a shortage of train units of a certain type, having different train types may lead to additional difficulties in dispatching.

Shunting Most of the shunting movements in and around stations result from the rolling stock schedule. These shunting movements have to be considered while planning the rolling stock, because they often cross many busy tracks in and around stations.

2.6.3 Crew Duties

A crew duty is the work that one crew member has to conduct on one day. In general, this is a sequence of train trips. Various characteristics of the crew duties have their influence on the delay propagation. Moreover, the train crew usually has its own opinion about the crew duties. This is an additional aspect which has to be taken into account in the crew duties. In 2000 and 2001, a conflict between the management of NS and the train crew, organized in different unions, led to a drop in punctuality, and finally the dismissal of the CEO and others (Nordbeck, 2003).

Crew transfers Each transfer within a crew duty is a possible propagation of a delay from one train to another. Therefore it would be advantageous for the punctuality to have as few crew transfers as possible: keep train and personnel together.

If transfers are planned despite this disadvantage, they should be long enough to absorb possible delays. However, this concept leads to long overlays of the personnel. Consequently, not all possible delays can be taken into account. A trade-off has to be made between possible delay propagation and inefficient duties.

Train team Both train drivers and conductors can transfer delays from one train to another. When these two crew members come from different trains, there are two sources of delay propagation. When the driver and the conductor have the same duty, i.e. they are a train team, there is only one possible source of disturbance. Therefore it is advantageous to keep the train crew members together.

Spare crew Because there will always be disturbances in train traffic it is often helpful to have additional personnel. These spare train drivers and conductors can be scheduled in real time to operate train services for which the originally planned personnel is not available. This can be caused by lateness or illness, but occurs more often in the case of large disturbances in the train traffic. However, having spare crew is costly. The spare crew often only works part of their duty, or even not at all. Again there is a trade-off, this time between the costs of spare duties, and the effectiveness of spare crew in case of disturbed train traffic, and the frequency in which these situations occur.

Crew satisfaction In general, crew members will deliver better work when they like their job. Many characteristics of the crew duties effect the crew satisfaction. The NS-project “Je bent erbij” (“You participate”) pointed at a long list of preferences for both train drivers and conductors (Abbink et al., 2004). Several of the positive characteristics for crew satisfaction are mentioned: many different train lines within a duty; different lines in a sequence of duties; a fair division of “nasty” work between crew depots; breaks halfway the duties; new rolling stock; few anticipated passenger aggression; intercity lines; no going up-and-down one line; and so on. Many of these preferences can have a negative influence on the punctuality or will cost additional duties. An additional problem is that every crew member has its own preferences.

2.7 Performance: Measurement and Perception

2.7.1 Reliability Measures

In the Netherlands, the position of all trains is registered in the TNV-database (Trein Nummer Volgsysteem = Train Number Tracing system). The train positions that are logged by the system are based on the signalling and interlocking system. Every passing time is automatically coupled with a train number and its scheduled time. To obtain arrival and departure times at stations, correction terms are applied (see Goverde, 2003).

- **Punctuality** Punctuality is probably the most widely used reliability measure in practice (Schaafsma, 2001), both in the Netherlands and abroad. This measure calculates the percentage of trains arriving within a certain number of minutes from the scheduled arrival time. In practice in the Netherlands, a three-minute margin is used. In most other countries, as well as for international comparisons, a five minute margin is more common.

  Besides arrival punctuality, also departure and start-up punctuality can be calculated. Note that, in order to calculate the punctuality figures, measuring points have to be chosen, usually a set of large stations. Computations of punctuality can also be made for a restricted set of trains; for example, long distance trains only.

  An improved punctuality measure would be the weighted punctuality, where each train is weighted according to its number of passengers. Note however that an estimate of the number of passengers has to be available in this case.

  For the formal punctuality figures in the Netherlands, those that are published and used in contracts with the government, not all trains and stations are considered. Only the arrivals at 34 large stations and only trains on the so-called *core network* are included, as is shown in Figure 2.7.

- **Transfer punctuality** Missing a connecting train often leads to long travel time extensions. With a frequency of twice per hour and a short transfer time, a small delay of the feeder train causes a delay of 30 minutes for the transferring passenger. For these situations, individual train delays give too little information. Therefore, the arrival times of feeder trains are compared to the departure times of the connecting trains. If the intermediate time is large enough for the passengers to transfer from the feeder train to the connecting train, then the transfer is maintained. The transfer punctuality is the percentage of transfers which is maintained during operations.

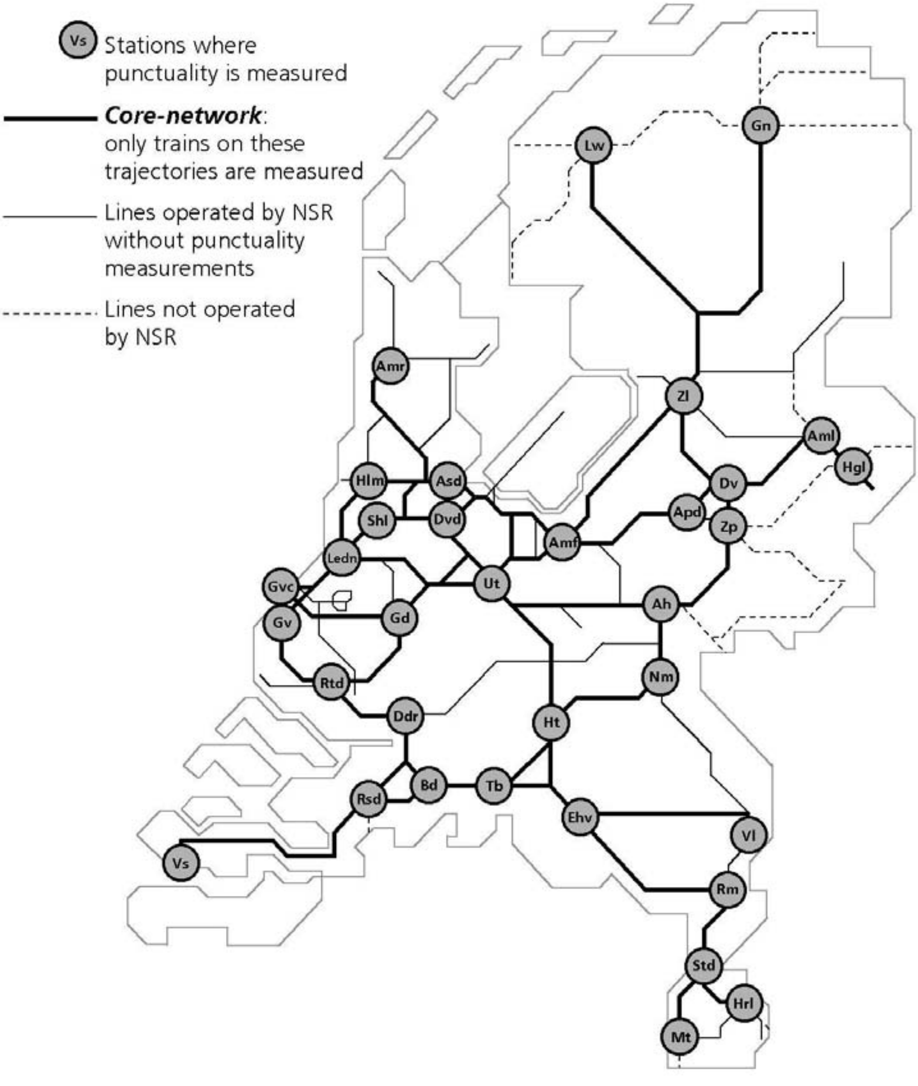


Figure 2.7: The network and stations which are included in the official punctuality figures.

Also this measure could be weighted, to represent the importance of the different connections in the measurements. Note that it is more difficult to count the number of transferring people for a certain connection, than just the number of passengers in a certain train.

- **Canceled trains** Canceled trains usually lead to large passenger delays, especially when frequencies are low.

  Note that canceled trains do not arrive and therefore never arrive on time, nor late: they are not included in the punctuality figures. Furthermore, it is hard to calculate the incurred delays of passengers. The next possible connection is not necessarily the same for all passengers of the canceled train. Furthermore additional transfers may be needed and, in the end, the delays of all the involved trains have to be taken into account.

- **Average train delay** The average train delay may be the most basic figure which can be applied. The disadvantage is that the few very large delays may have a strong influence on the average delay.

- **Average passenger delay** The average passenger delay represents what the average passenger experiences. However, to attain this figure, figures on train delays and canceled trains, as well as on the numbers of passengers and transfers have to be available.

All these measures have their advantages and disadvantages. Note that none of the discussed measures includes the variation in day-to-day operations.

Goverde et al. (2001) state that the arrival and departure times, which are derived from the TNV-data are not accurate. This has its origin in the fixed correction terms, which 'define' the running time between the last measuring point before the platform and the platform and between the platform and the first measuring point thereafter. Obviously, these fixed running times are not always adhered to. To acquire more accurate arrival and departure times Goverde et al. (2001) developed TNV-prepare. This system uses the passage times of subsections, which are not directly coupled to train numbers. Starting with passage times at section ends which are coupled to train numbers and then following the train subsection-by-subsection, makes it possible to estimate the departure and arrival times more accurately.

Due to a review of the fixed correction terms, the TNV-data are somewhat more accurate now than in 2001.

The reasons for disturbances are registered by the Traffic Control Center. Also this registration is not satisfying. First of all, only a small portion of the disturbances is administered. Secondly, it is not always easy to verify which disturbance caused a certain delay. Therefore the Ministry of Traffic and Waterworks, ProRail, NS and Railion intend to set up a new registration system (Schulz van Haegen-Maas Geesteranus, 2004).

2.7.2 Other Quality Characteristics

Reliability is one of the key factors in the success of public transport. However, many more characteristics determine the overall quality. Although most of the items below fall outside the scope of the thesis, some of them are influenced by actions taken to improve reliability.

- **Travel time** Travel time is one of the prime determinants of travel mode choice: time is money. It is noteworthy that travel time, in case of public transport, includes waiting times, transfer times, and the time from the origin to the access point of the public transport system, and from the public transport's exit point to the destination. When traveling by public transport, someone's opinion will be more positive when travel times are smaller. Especially the travel time ratio between the different travel modes is important.

- **Direct connection** A direct train connection has several advantages for passengers. First of all, time is needed to go from the feeder train to the connecting train. This connection time has to be added to the travel time. Next, waiting times are perceived as worse than travel times (Rietveld et al., 1998). In case of a two-train trip, there are two waiting times instead of one for the direct connection. Furthermore, because of disturbances, there is always a probability of missing a connection, which leads, especially in case of low frequencies, to large delays.

- **Price** The prices in public transport have a large influence on the choice of a transport mode. Every time that tariffs are changed, passenger organizations and passengers are sceptical. Often prices in public transport are bounded or even determined by the government.

- **Frequency** The timetable has to match the desired travel moments of the passengers. But because all passengers are different, not all their desires can be met by a few trains on a certain route. When there are more trains running on this route, the probability that an appropriate departure (or arrival) time is available for the passengers increases. Except for some peripheral lines, the frequencies in the Netherlands are usually so high that there is always a train with a planned arrival time around the desired arrival time.

When trains run so often that there will always be a train soon, passengers will stop using the timetable. Such a high frequency can also have other implications. For example, the operator may decide to communicate that there will be a train every 5 minutes. Furthermore, passenger connections do not have to be taken into account anymore in timetabling.

- **Travel information** The accessibility of public transport partially depends on the travel information provided to the public. When a potential passenger is not able to acquire departure and arrival times, he may decide to use his car. Therefore, pre-trip travel information is not only available at stations, but also accessible in several other ways. Examples of sources of pre-trip information at home are a timetable booklet, the internet, and a telephone-service. At the station, information can be acquired at the ticket counters, from the departure tables, from large automatic displays with the oncoming departures, from displays at platforms with the first departing train from that platform, and from automatic signs on the trains. Note that some of the described devices are also used for real-time travel information, as is explained below.

- **Information in case of delays** Especially in case of delays, information to travelers is important, both to inform the passengers about the delays or cancelations and to inform them about alternative travel options. Information should be available at home, at the station and in the train. Furthermore it always has to be up-to-date, and cover all traveling options. In the Netherlands, limited real-time information is available via internet. At stations several types of displays are used to inform the passengers about delays and changes in departure platforms. Furthermore, service personnel keeps the passengers informed.

- **Seating probability** A comfortable journey increases the attractiveness of the train. Especially for long trips it is important that all passengers can be seated. A seating guarantee can only be given with a seat reservation system.

- **Safety** The safety of a train trip has two main aspects: the safety of the railway system with respect to accidents, and the social safety at stations and in trains.

- **Cleanliness** Clean trains and stations increase the traveling pleasure and comfort. A clean environment also increases the feeling of safety.

- **Pre and post transport** The door-to-door trip of a train passenger often includes other modes of public transport, such as bus or metro. The quality of these modes does not only influence the perception of the train traffic, but also the number of passengers.

- **Facilities** Facilities at stations also have a positive influence on the attractiveness of railway traffic. Some of these facilities are really part of going from A to B (bicycle-sheds and waiting rooms), others make the journey more enjoyable (bookshops), and yet others have no direct connection to the train trip, but are easy to use for train passengers (catering and grocery stores).

2.7.3 Perception

The public perception of the railway product is subjective and depends on many factors. Not only the travelers' train trips influence their perception, but also the trips to and from the train station. Furthermore, stories from colleagues and articles in newspapers affect someone's judgement. When background music is played at the station waiting may seem shorter.

In general, people do not like uncertainty. Therefore it is also important to inform passengers adequately when delays occur. Not only to tell them which train is delayed or canceled, but mainly to tell them what alternative travel options they have. Furthermore delays are less annoying when it is made clear why a train is late. In the car the traveler is in charge, in the train he or she is not, and therefore he or she can blame the railway system.

Many of the travel characteristics in Section 2.7.2 are of a subjective nature. Even the quantifiable delay measures in Section 2.7.1 cannot exactly quantify the perception of the passengers. To attain insight in the passenger perception of the offered train services, NS, the government and passenger organizations support an independent passenger satisfaction inquiry which is held every three months (NS, 2004b).

The story below illustrates that reliability measures are not the only factor determining the passengers' perception.

The storm of October 27th, 2002 At the end of October 2002, a large storm crossed the Netherlands and other parts of Europe. Due to the date in early fall, all trees released all their leafs at the same day. Moreover, the leafs were still relatively fresh. Together with the moist, this led to greasy and slippery tracks.

Early on the specific day, Sunday the 27th, hardly any train ran on time: on a 3-minute basis, the punctuality was 38.5%. Trains were slipping and sliding along the tracks and the overhead power lines were torn apart on several lines. In the early afternoon, it was decided to suspend all railway traffic for the rest of the day to prevent more damage, leading to 67.2% of the trains to be canceled that day.

The following Monday started in chaos, because train units were not where they were supposed to be and many train units even had so called *square wheels*: flat sides because of slipping on the tracks, mainly at departure and arrival at stations.

The number of trains in repair was so high that NS had to decide to cancel some of their train services for several weeks. The remaining trains were overcrowded, and still had problems with the slippery tracks. November 2002 produced a punctuality of 78.5%. The percentage of canceled trains however was relatively low, due to the fact that cancelation is registered with respect to the timetable which is communicated to traffic control 36 hours in advance. So, only the first few days after the storm showed high cancelation rates.

In an attempt to improve reliability, NS spent €18 million on *fall measures* in 2003. Due to these investments and the absence of extreme weather, performance was much better in November 2003. At least, that was the public opinion. NS did not have to cancel many trains and seating capacity was sufficient.

However, the performance measures showed other figures: punctuality was 74.0% in November 2003, 4.5%-point less than in November 2002, due to many moderate delays (3 to 8 minutes). Furthermore, due to the 36-hour procedure, the planned cancelation of 6% of the trains in November 2002 cannot be found back in the official figures.

The situation described here shows how delicate performance measures are. Although the official performance measures showed a quality decline, client satisfaction research showed an increased overall judgement for the fourth quarter of 2003 with respect to one year before (from 6.3 to 6.6 on a scale from 1 to 10). Even the punctuality of trains was perceived better (from 5.2 in the fourth quarter of 2002 to 5.7 in 2003 (NS, 2004b)).

2.8 A Literature Review

Over the last twenty years, a range of researchers has studied railway timetabling and punctuality issues. Literature reports on different types of railway timetabling research. The overview in this section starts with selected literature that gives an overview of railway practice and research. The main part of this review focusses on timetable construction and evaluation. This part start with timetable construction. Then analytical delay models are discussed, starting with max-plus algebra, followed by stochastic models. Finally, the focus is on railway simulation.

2.8.1 Railway Planning in General

A broad insight in railway systems is provided by Pachl (2002). He starts with the introduction of railway terms. Then he gives an overview of how railway systems work. The second part is more based on modeling railway systems. To that respect he discusses capacity research, scheduling problems and control of the operations.

Assad (1980) gives an extended overview of railway modeling until 1980. He discusses a broad range of railway planning literature. The article mainly focusses on cargo. Cordeau et al. (1998) provide a more recent overview of optimization and planning models for many problems in both passenger and cargo railway problems. Schwanhäußer (1994) discusses the status of German railway research. Amongst others, he discusses models on capacity, timetabling models and timetable evaluation based on simulation.

Huisman et al. (2005) provide an overview of the use of operations research in passenger railway transportation. Besides an overview of international literature on this subject, they pay extra attention to models that are in use by NS. Most of these models are clarified by examples. Kroon (2001) also focusses on the use of operations research models in the planning of NS. Each of the models is illustrated with a practical example.

2.8.2 Railway Line Planning and Timetabling

Given the expected demand for railway transport, the railway operator can start its planning process. For the planning on a tactical level (say 6 to 18 months), the physical infrastructure will be given. However, it should be noted that this infrastructure is the result of earlier, strategic studies. On a societal level, these studies also include other modes of (public) transport. For example, Van Nes (1997) discusses hierarchical public transport networks in a multimodal environment. He also presents a model that optimizes the line spacing and stop spacing for the network of one mode. Earlier literature on multimodal network design is also described.

This thesis focusses on the tactical planning level. The first step on this level is line planning. An extensive description of the line planning problem is given by Goossens (2004). He describes several models and algorithms to solve these. His models account for different service types (local trains, intercity trains, ...), direct connections, seating capacity, operating costs and other timetable qualities. Further literature related to line planning is also described by Goossens (2004).

When the line plan is known, timetabling can commence. Serafini and Ukovich (1989) developed a mathematical model for the PESP (Periodic Event Scheduling Problem). Using this model, timetabling restrictions can be formulated and, if there is any, a feasible cyclic timetable can be found. Based on PESP, Schrijver and Steenbeek (1994) developed the CADANS module for DONS (see also Section 3.1). DONS is short for Designer Of Network Schedules and is the graphical user-interface for the semi-automatic timetabling system CADANS used by NS and ProRail. In DONS, train lines can be defined, including rolling stock and passenger connections and other characteristics. The CADANS-module then checks whether a timetable respecting all these restrictions is possible. If so, a possible solution is presented (see also Hooghiemstra, 1994).

Assuming certain preferences for the timetable features, some optimization models have been proposed. Nachtigall (1998) extensively studied periodic timetabling models based on PESP. Nachtigall and Voget (1997), use it to minimize waiting times for passengers. Furthermore, Nachtigall (1996) focusses on networks with different line frequencies.

Models by Goverde (1999) and Peeters and Kroon (2001) optimize characteristics like layover times, passenger connection times and inter train distances. Kroon and Peeters (2003) also describe the use of variable running times within the PESP model. Liebchen (2003) extends the PESP model with symmetry constraints. Although this often leads to suboptimal solutions, these constraints can speed up the process of finding a good solution considerably.

Peeters (2003) provides an extended description of the PESP as a generalization of the conventional Cyclic Railway Timetabling Problem. He also reformulates the problem as the Cyclic Periodicity Formulation (CPF), based on tensions. He shows that the CPF performs better on several real life cases.

Caprara et al. (2002) solve the timetabling problem for one line, and optimize with respect to preferred departure and arrival times and other timetable characteristics. The problem is represented by a directed multigraph. Lagrangian relaxation is used to solve the integer linear programming model derived from the graph.

Although most models use time distances to model the safety distance between trains, most signaling systems in Europe are based on block occupation. Wendler (1995) describes for several safety systems how these block occupations can be modeled.

Gröger (2004) describes how asynchronous simulation is used to construct timetables based on block occupation rather than time distances. Special attention is directed at the conflict resolution of two or more train paths.

In addition to these models that mainly focus on open tracks, several models exist for platform assignment and routing of trains in stations. Zwaneveld et al. (1996b) formulate this problem based on the Node Packing Problem. Carey and Carville (2003) developed constraints and objectives for routing trains in stations. The used heuristics are based on 'manual' methods applied by planners. Billionnet (2003) describes the platforming problem as a Graph Coloring Problem. He formulates two different integer linear programs which can be used to solve the problem. The first formulation is a more intuitive formulation, the second one provides better computational results.

2.8.3 Railway Reliability, Analyzing Realization Data

Analyzing realization data is quite practical in nature. In many cases locations or train lines with a high unpunctuality level are studied to find bottle-necks. This kind of analysis is often carried out by railway companies themselves. The results often remain hidden for the scientific world.

A more theoretical study of realization data is performed by Goverde et al. (2001), who analyzed realization data for one week for trains around Eindhoven, an important station in the southeast of the Netherlands. They searched for the distributions of arrival and departure times as well as of dwell time elongations. Yuan and Hansen (2004) analyzed the train traffic at and around The Hague HS, a crowded station in the west of the Netherlands, for the same purpose.

Bruinsma et al. (1998) estimate the unreliability in public transport chains. They use gamma, lognormal, and Weibull distributions to fit realization data of different public transport modes. They conclude that much of the unreliability is due to missed connections. In a later paper they also describe options to decrease the unreliability. These include the use of the bicycle as entrance mode, longer planned connection times, and a system to prevent bus drivers from departing early (Rietveld et al., 2001). In the latter paper they use the 64/27 ratio for a minute of unreliability compared to a minute of planned travel time, which they found earlier (Rietveld et al., 1998).

A Norwegian study by Olsson and Haugland (2004) studies the correlation between the arrival punctuality on one hand and several factors on the other hand, such as the departure punctuality, the number of passengers, the number of passengers per seat, and the capacity consumption of the infrastructure. Despite the correlations that were found, no timetable improvement strategies were developed.

2.8.4 Railway Reliability, Analytical Models

Max-plus algebra is an analytical approach for evaluating timetables on robustness. Some key characteristics, like minimal cycle times, are easily calculated with max-plus algebra (Subiono, 2000, Goverde and Soto y Koelemeijer, 2000, Van Egmond, 1999, De Kort, 2000). PETER, based on max-plus algebra, is a performance evaluator for timetables (Soto y Koelemeijer et al., 2000, Goverde and Odijk, 2002). Current max-plus research in the field of railways focuses on the inclusion of stochastic disturbances in the models. Hansen (2000) uses both queuing theory and max-plus algebra to study the capacity and stability of train movements, but only in stations. Another probabilistic approach by De Kort et al. (2003) uses a max-plus model to assess the capacity of the railway infrastructure. This probabilistic capacity assessment is based on unspecific timetables. This means that train lines are known, but there are no actual departure and arrival times available. For a more detailed description of max-plus algebra and its possibilities, see Section 3.2.

Weigand (1981) develops a model that is able to evaluate delay propagation of exponential disturbances throughout a railway network. He also shows how to determine the minimal amount of running time supplements that is necessary to achieve a certain average arrival delay.

Wakob (1985) introduces a queueing model to assess the capacity for a given subnetwork. His method is known as Wakob's razor, and is based on random train arrivals without a timetable. Schwanhäußer (1994) describes the use of queueing models for determining the expected value of anticipated delays and other timetable characteristics. These queueing models are based on train frequencies and running times only, not on timetables. Therefore they are better suited for a capacity analysis of the infrastructure than for timetable evaluation, let alone for timetable optimization.

Huisman and Boucherie (2001) and Huisman et al. (2002) developed a stochastic analytical waiting time model for analyzing delays at a double track section. Huisman (2005) is also able to apply these queueing models to scheduled timetables. With his models he also analyzes the consequences of several timetable characteristics for the delay propagation.

Higgins et al. (1995) come up with a model to quantify so called risk delays for a single track line. Higgins and Kozan (1998) also developed an analytical model to quantify the expected delays of individual passenger trains in an urban rail network.

Carey and Kwieciński (1995) mainly focus on recovery times in their stochastic analysis. Carey (1999) also uses heuristic measures for timetable reliability and includes behavioral response (Carey, 1998).

2.8.5 Railway Reliability, Simulation

Other researchers use simulation as a tool to analyze the influence of delays on the train circulation, given some traffic scenario. SIMON is a Swedish software tool using simulation of the whole network (Wahlborg, 1996, Bergmark, 1996). Amongst others, VirtuOS (König, 2001) and SABINE (Fischer et al., 1995) are used in Germany, and Open Track (Hürlimann, 2001) is a railway simulation program developed at ETH Zürich. UX-SIMU is used for detailed simulation of railway traffic in Denmark (Kaas, 2000).

The literature above is mainly about the simulation software itself and sometimes on a simple comparison of multiple timetables. More thorough research of the impact of timetabling principles on reliability based on simulation is hardly found.

However, Carey and Carville (2000) focus on the evaluation of delay propagation in large stations by simulation, and on the improvement of the underlying timetables (Carey and Carville, 2003).

Rudolph (2004) develops a strategy to improve the running time supplement allocation. She shows that the reallocation of the proportionally distributed supplements to locations just before major railway nodes decreases the average delay. This strategy is based on a theoretical analysis of the timetable and operations. The described method of moving supplements does not lead to a different timetable, at least not for the large nodes and stations. However, Rudolph proposes to move the supplements to the dwell time in the internal plan. This leads to a difference between the internal operator's schedule and the published timetable. In fact, this distinction is in line with earlier research by Schaafsma (2001) (see Section 3.3.4). Rudolph uses detailed simulation to verify the practical applicability of this strategy.

Rudolph and Demitz (2003) and Demitz et al. (2004) describe the improvement of the timetable for North Rhine Westphalia by use of RailSys. RailSys is used for the (non-automated) construction of timetables based on block occupation, and detailed simulation.

Furthermore, Middelkoop and Bouwman (2000, 2001) describe the use of SI-

MONE for the evaluation of several traffic scenarios in the Netherlands. SIMONE is capable of simulating the entire Dutch railway network. Although most SIMONE studies at NS and ProRail are of a comparative nature, some of them analyze the influence of planning norms on the punctuality. This is described by for example Al-Ibrahim and Oldenziel (2004), and De Klerck-Salm (2004).

Wojtkowski (2004) describes an iterative process of simulation, bottle-neck analysis and timetable adjustments to find a better timetable. Although his method is quite cumbersome, he realizes significant punctuality gains in his simulations.

SIMONE is used later in this thesis to execute theoretical analyses on the basis of simulation. Further details are provided in Section 3.3.

Chapter 3

Timetabling, Max-Plus Algebra, and Simulation

The basic ideas of timetabling are discussed in Chapter 2. In this chapter some of those ideas are formalized. First a mathematical model is formulated for the construction of timetables. Later, two distinct timetable evaluation methods are discussed: max-plus algebra and simulation.

This chapter starts with the introduction of a mathematical model for cyclic timetabling. In particular, the timetabling system DONS is described in Section 3.1. Many of the timetables used later on in this thesis have been created with DONS.

In Section 2.8, max-plus algebra was mentioned as a possible method for timetable evaluation. In particular, the software tool PETER can evaluate DONS-timetables, using max-plus algebra. The principles of this evaluation method are described in Section 3.2.

Another wide-spread method for the evaluation of railway systems is simulation. Section 3.3 starts with a short discussion on simulation in general. The main part of this section is devoted to the simulation tool SIMONE, which is used throughout the remainder of this thesis for evaluation purposes. SIMONE is linked to the DONS timetabling system to facilitate the evaluation of timetables which are constructed with DONS.

3.1 Timetabling and DONS

Constructing a timetable is a very complicated matter with a large number of interrelationships. These relations include infrastructural and safety constraints, passenger connections, and layover times. Due to these relations, timetable modifications on a peripheral line can have severe impacts on the complete network. Especially at and around large nodes, these relations become very disorderly. This is mainly due to the fact that choices for the timetable made at one place influence the timetable in other parts of the network.

The structure and complexities of timetables are presented in this section on the basis of the software program DONS: Designer Of Network Schedules. DONS (Hooghiemstra, 1996, Hooghiemstra et al., 1999) was developed for NS and ProRail for semi-automatic timetable construction. By supplying this system with the necessary information and preferences, DONS will look at the problem on a network wide scale.

Important to note is the fact that DONS is not an optimization tool: it either provides a feasible timetable or it tells that the user defined restrictions create an infeasible problem. When a feasible timetable is found, this timetable can be post-optimized. The post-optimization process can only shift arrival and departure times a little bit, but the train orders are fixed. When the problem is infeasible, DONS provides some information on conflicting restrictions. In Chapter 6 we develop a new stochastic model which does optimize the timetable with respect to delays.

Cyclic timetables An important principle used by DONS is the cyclicity of the timetable. This means that the timetable may repeat itself over and over again.

The cyclic, or periodic, nature of this problem makes it equivalent to the Periodic Event Scheduling Problem (PESP) described by Serafini and Ukovich (1989). DONS and its solver CADANS are based on the PESP. Some other timetabling models were mentioned in Section 2.8.

DONS can handle cycle times of 60 and 120 minutes, which means that all trains run every hour or every two hours respectively. It is relatively straightforward to adopt the software for other cycle times. DONS is used in later chapters of this thesis to construct cyclic timetables.

3.1.1 The DONS Modules

The DONS system consists of four integrated parts: a graphical user interface, a database, and the solvers CADANS and STATIONS.

- **Database** The DONS database contains all information provided and saved by the users for the definition of the problem, as well as the computed timetables. The database is also the part of DONS which is able to communicate with other software. For example, the simulation tool SIMONE receives most of its input from the DONS database.

- **Graphical User Interface** The Graphical User Interface (GUI) provides a user friendly graphical interface. With graphical representations of the network and the nodes, and a tabular representation of many timetable characteristics, it is easy to enter or adjust the constraints for the desired timetable, which are stored in the database. After a feasible solution has been constructed by DONS, the GUI can also present these results in the form of a time-distance diagram or a platform occupation chart.

- **CADANS-solver** When instances are created with the aid of the GUI and stored in the database, the solvers can be used to create a One-Hour Timetable or a Platform Occupation Chart. The CADANS module (Schrijver and Steenbeek, 1994) is used to create a timetable with feasible arrival and departure times for all timetable points. Although CADANS does not take into account what happens inside a station, conflicting train routes can be specified. To run CADANS, detailed infrastructure information for the nodes is not necessary; defining the tracks that enter and leave a timetable point suffices. More about the restrictions and the modeling in CADANS can be found in Section 3.1.2.

- **STATIONS-solver** The STATIONS module (Zwaneveld et al., 1996a,b), as the name gives away, is used to produce plans for the nodes. Given a network wide CADANS-solution, STATIONS searches for a feasible platform assignment at a station, and a feasible routing between the open tracks and the platforms. STATIONS constructs a feasible solution if possible, or it indicates that, given the user's preferences and the CADANS-solution, no feasible platform assignment or routing is available. A detailed track and platform layout of the considered station have to be available for this module.

3.1.2 CADANS: Constraints, Modeling and Solving

CADANS is able to tell the DONS-user whether the provided constraints can be combined into a feasible timetable. In case of feasibility, CADANS will also provide such a timetable.

Decision variables The timetable which has to be constructed is made up of departure and arrival times at timetable points. By means of distances and speed characteristics of the chosen rolling stock and running time supplement settings, running times are calculated. Although the running times can be adjusted manually, they are provided to CADANS as fixed numbers. Therefore it is easy to derive arrival times when departure times are known. Consequently, it suffices to only have the departure times as decision variables.

Departure and arrival times are integers in the cycle interval. From here on, let us assume that DONS is used to develop a One-Hour Timetable with a one-minute precision and a cycle time of 60 minutes. This gives the departure time

$$d_n^t \in [0, 59], \tag{3.1}$$

with t the number of the train line, completed with a digit for direction and possibly a sequence number when running more than once per cycle, and n the timetable point or station.

Constraints There are several types of constraints in timetabling. The first type consists of *process* constraints, which are the restrictions on running times and dwell times of the defined train lines. Secondly there are *safety* constraints. These make sure that there is at least a certain time distance between the trains. Then there are *market* constraints which emerge from passenger preferences. Finally, there are *logistic* constraints which formulate the operator's desire to construct an efficiently operable timetable.

The process constraints define the running times and dwell times. Running times are always fixed, but dwell times can either be fixed or chosen in some interval. Constraints on the running times themselves are only needed for the calculation of the arrival times afterwards. When, for train line t, n' is the timetable point following timetable point n, a running time constraint would look like

$$a_{n'}^t - d_n^t = r_{n,n'}^t \text{ modulo } T, \tag{3.2}$$

where $r_{n,n'}^t$ is the fixed running time and T is the cycle time. The *modulo* T is included to ensure the validity of the constraint when the departure time is in an earlier cycle than the arrival time.

Equation 3.2 can also be written as

$$a_{n'}^t - d_n^t = r_{n,n'}^t + p \cdot T, \quad p \in \mathbb{Z}. \tag{3.3}$$

An example for a one-hour timetable ($T = 60$), and a running time of 25 minutes is a departure at .47 every hour and an arrival at .12. Simply subtracting the arrival time from the departure time would give $12 - 47 = -35$, while the running time is 25. So the equation above corrects for this: $-35 + 1 \cdot 60 = 25$.

In the remainder of this thesis, the following notation will be used,

$$a^t_{n'} - d^t_n = [r^t_{n,n'}]_T. \tag{3.4}$$

Still, as reasoned above, these restrictions have not been incorporated in the mathematical model. The fixed running times are used to construct constraints, which are formulated in such a way, that only departure times are used as variables.

The other type of process constraints are the dwell time constraints. When the dwell time $s^t_{n'}$ is fixed, these constraints look like

$$d^t_{n'} - d^t_n = [r^t_{n,n'} + s^t_{n'}]_T, \tag{3.5}$$

or, when the dwell time has to be at least $ms^t_{n'}$ and at most $Ms^t_{n'}$:

$$d^t_{n'} - d^t_n \in r^t_{n,n'} + [ms^t_{n'}, Ms^t_{n'}]_T. \tag{3.6}$$

An example will be given for the 3000-line, an intercity line running from Den Helder via Amsterdam, Utrecht (Ut), Ede-Wageningen (Ed), and Arnhem (Ah) to Nijmegen. The 3000-line runs twice per hour, leading to a 3001 and 3003 train in the forward direction (from Den Helder to Nijmegen), and a 3002 and 3004 train in the opposite direction (from Nijmegen to Den Helder).

Assume that the running time between Utrecht and Ede-Wageningen is 21 minutes and between Ede-Wageningen and Arnhem 9 minutes. Now, if the dwell time in Ede-Wageningen should be 1 minute, this generates the following restriction:

$$d^{3001}_{Ed} - d^{3001}_{Ut} = [r^{3001}_{Ut,Ed} + s^{3001}_{Ed}]_T = [22]_{60}. \tag{3.7}$$

When the dwell time in Arnhem should be at least 2 minutes and at most 5 minutes this gives

$$d^{3001}_{Ah} - d^{3001}_{Ed} \in \left[r^{3001}_{Ed,Ah} + [ms^{3001}_{Ah}, Ms^{3001}_{Ah}]\right]_T = [11, 14]_{60}. \tag{3.8}$$

Safety constraints, which can also be viewed as infrastructural constraints, make sure that two trains do not use the same infrastructure at the same time. When the departure of train t' from station n' is not allowed to be in between l minutes and u minutes after the departure of train t from station n, the following constraint can be constructed:

$$d^t_n - d^{t'}_{n'} \notin (l, u)_T. \tag{3.9}$$

Note that the use of parenthesis '( )', as opposed to brackets '[ ]', indicates that a time difference of exactly l minutes or exactly u minutes is allowed.

The time difference required between two train departures, and the resulting lower and upper bounds l and u, can result from quite different situations, as can be seen from the following examples. The nodes n and n' can be the same timetable point.

First, to obtain inclusion intervals for all constraints, we rewrite (3.9) as:

$$d_n^t - d_{n'}^{t'} \in [u, l + T]_T. \tag{3.10}$$

Besides the 3001-train, there is also a local 7501-train from Utrecht to Ede-Wageningen. These two train lines share the same infrastructure, for which the safety system requires a 3-minute distance between each pair of trains. This leads to

$$d_{Ut}^{7501} - d_{Ut}^{3001} \in [3, 57]_{60}. \tag{3.11}$$

A comparable safety constraint applies to the arrival in Ede-Wageningen. The running time from Utrecht to Ede-Wageningen is 22 minutes for the 3001-train and 30 minutes for the 7501-train, resulting in an 8-minute difference:

$$d_{Ut}^{7501} - d_{Ut}^{3001} \in [3 + 8, 57 + 8]_{60} = [11, 65]_{60}. \tag{3.12}$$

However, this could lead to a conflicting situation: for example $d_{Ut}^{7501} = 0$ and $d_{Ut}^{3001} = 4$ satisfy both restrictions. These departure times from Utrecht would imply that the 3001-train overtakes the 7501-train, which is impossible on the available infrastructure. To prevent this from happening, the two restrictions are combined to

$$d_{Ut}^{7501} - d_{Ut}^{3001} \in [3 + 8, 57]_{60} = [11, 57]_{60}. \tag{3.13}$$

In general terms, this can be written as

$$d_n^t - d_n^{t'} \in [u + (\rho_{n,n'}^{t,t'})^+, l + (\rho_{n,n'}^{t,t'})^- + T]_T, \tag{3.14}$$

where $\rho_{n,n'}^{t,t'}$ is the running time difference of trains t and t' between n and n' (with $\rho_{n,n'}^{t,t'} < 0$ when train t is faster than train t'). Furthermore, $(\rho_{n,n'}^{t,t'})^+$ and $(\rho_{n,n'}^{t,t'})^-$ represent $\max(\rho_{n,n'}^{t,t'}, 0)$ and $\min(\rho_{n,n'}^{t,t'}, 0)$, respectively.

The market constraints represent passenger preferences. This includes for example passenger transfers and an even distribution of trains over the hour. It would be preferable to run the 3001 and 3003 in approximately a 30-minute frequency:

$$d_{Hdr}^{3003} - d_{Hdr}^{3001} \in [29, 31]_{60} \tag{3.15}$$

and likewise for the opposite direction.

When dwell times are not fixed, the total time to reach Nijmegen from Den Helder can differ several minutes between the 3001 and the 3003. To avoid this problem, the same constraints should be added for other departures on the line.

For passenger transfers it is important that passengers do not have to wait too long between arrival and departure. On the other hand they need time to alight their feeder train, be able to go from the arrival platform to the departure platform, and board the connecting train.

For example, passengers arriving with the intercity train from Amsterdam in Utrecht (train 3001) want to transfer to the local service towards Ede-Wageningen and Arnhem (train 7501). The transfer time should be at least 4 minutes and at most 7 minutes. The travel time from Amsterdam (Asd) to Utrecht is 28 minutes.

$$d_{Ut}^{7501} - d_{Asd}^{3001} \in [32, 35]_{60}. \tag{3.16}$$

But as we have seen before, the 3000-line runs twice per hour. Here we assume that it runs exactly every 30 minutes. For train 7501 it does not make any difference to which of the two trains in the hour it connects, the 3001 or the 3003. This means that the 7501 leaves Utrecht either 4 to 7 minutes after arrival of the 3001 in Utrecht, or 4 to 7 minutes after the arrival of the 3003 in Utrecht. Assuming that trains 3001 and 3003 are exactly 30 minutes apart, and because of the cyclicity, this can be modeled as the intersection of two larger intervals, $[32-30, 35]_{60}$ and $[32, 35+30]_{60}$:

$$d_{Ut}^{7501} - d_{Asd}^{3001} \in [2, 35]_{60}, \tag{3.17}$$

$$d_{Ut}^{7501} - d_{Asd}^{3001} \in [32, 65]_{60}. \tag{3.18}$$

When train 7503 also exists, we still need to model the connection in the same way. That is because it is unsure whether train 3001 connects to the 7501 and the 3003 to the 7503, or the other way around.

Besides desires of the passengers, logistic constraints can be added to make the timetable more efficient for the operator. One possibility consists of short layover times of rolling stock at the end of the train lines. There is a minimum time required for these layovers to enable personnel to prepare the train for running in the opposite direction. When the running time of the 3000-line from Arnhem to Nijmegen is 12 minutes and the layover time has to be at least 4 minutes, but at most 12 because of efficiency, this would give

$$d_{Nm}^{3002} - d_{Ah}^{3001} \in [16, 24]_{60}. \tag{3.19}$$

Furthermore, absolute departure times can be defined. This option is mainly used for international trains. For example, the ICE-train from Frankfurt to Amsterdam (train 122) passes the border at Zevenaar-grens (Zvg) at 36 minutes past the hour:

$$d_{Zvg}^{122} = 36. \tag{3.20}$$

Note that the modulo 60 is not necessary in this case. To represent all constraints as the difference between two departure times the dummy departure $d^0 \equiv 0$ can be introduced, and we have

$$d_{Zvg}^{122} - d^0 = 36. \tag{3.21}$$

Solving CADANS does not optimize the timetabling problem. In case the problem is solvable, CADANS will only provide a feasible solution. Otherwise it will produce an indication which constraints are too tight.

CADANS is based on constraint programming (Schrijver and Steenbeek, 1994). The different types of constraints (described above) define the restrictions which a solution has to satisfy. A feasible solution is a set of departure times d_n^t which satisfy all constraints. Every integer τ, with $0 \leq \tau \leq 59$, is called a *clock value*, and a set of clock values is called a *clock set*. The difference between two clock values is always computed modulo 60, unless indicated otherwise.

The input for CADANS is the set of all preferences, as explained earlier in this section. CADANS translates these preferences into constraints. The constraints are represented by a constraint graph, where the nodes represent the decision variables, or departure times, and the directed edges represent the constraints, or time differences. More precisely, the edge values represent the lower and upper bounds l and u.

Several edges between the same pair of nodes are combined into one edge if possible. When two nodes are connected by an edge that only allows one unique value, one of both nodes and the edge can be removed. The edges to this node are now reconnected to the other node, while the edge values are adjusted for the value of the removed edge.

Combining all paths between two nodes often leads to tighter constraints. It may even lead to conflicts. In other words: the problem may appear to be infeasible.

In the beginning of the solution process, every decision variable d_n^t has its own clock set $D_n^t = \{0...59\}$. If no absolute departure times have been defined (see equation (3.20)), then one of the departure times can be fixed; i.e, its clock set will have only one element. This will not have any consequences for the feasibility, because without absolute times the whole solution can be shifted by an integer between 0 and 59, modulo 60.

Starting from the fixed departure time, or from an earlier defined absolute departure time, the constraint edges will be used to tighten the clock sets. This is called constraint propagation.

Then a tight clock set D_n^t is chosen and its decision variable d_n^t is fixed to one of the values in the set, $\tau \in D_n^t$. Again, the clock sets can be tightened. When one of the clock sets is empty, the fixation leads to infeasibility and τ will be removed from the clock set. All other clock sets are reset to their size before fixing d_n^t to τ. Then the same d_n^t is fixed to another $\tau' \in D_n^t$, or another $d_{n'}^{t'}$ is fixed and the procedure starts again.

In the case that a fixation does not lead to empty clock sets, a next departure time can be fixed. When all departure times have been fixed in this way, a feasible timetable has been found.

When a fixed departure time $d_{n''}^{t''}$ leads to an infeasibility, while $d_{n''}^{t''}$ was the last element of the clock set $D_{n''}^{t''}$, one of the other fixations has to be undone. An intelligent way of back-tracking is applied. If there is no other fixed departure time, the problem is infeasible. This means that no timetable exists which satisfies all restrictions. In that case, CADANS does provide the user with a set of specified constraints which are in conflict with each other. Then it is up to the user to relax one or more of these, or possibly other, constraints, and CADANS can be run again.

STATIONS Although many relations between trains in and around stations can be defined within the network environment of DONS, detailed modeling and analysis of the stations does not fit in CADANS. The STATIONS module within DONS is able to assign trains to platforms, and to route trains through stations. The simulation tool SIMONE (see Section 3.3) only considers the modeling of the open tracks, corresponding to the CADANS part of DONS. The simulation model FRISO (see Section 3.3.4) is currently (2005) in development for more detailed simulation of station areas and smaller parts of the network.

3.2 Max-Plus Algebra for Timetable Evaluation

Many events, in railway and other systems, cannot take place before the preceding event has finished. In many cases, even more than one preceding process must have been finished before a certain event can take place. In that case, the event cannot start before the latest, or maximum, completion time of its predecessors.

In mathematics, such problems can be represented by for example precedence graphs or Petri nets. When a time factor is introduced into a Petri net, it is called a timed event graph. To facilitate calculations for these graphs, a new kind of algebra has been developed: max-plus algebra. A timetable can be represented as a timed event graph and therefore also in max-plus algebra. Using this algebra, several characteristics of the timetable can be calculated.

A thorough description of precedence graphs, Petri nets, timed event graphs and max-plus algebra is given by Baccelli et al. (1992).

3.2.1 Principles of Max-Plus Algebra

Max-plus algebra only uses two operators: taking the maximum and addition. The o-plus symbol, $\oplus$, is used for *taking the maximum* and the o-times symbol, $\otimes$, is used for *addition*. For example, $a \oplus b$ and $c \otimes d$ in max-plus algebra, refer to $max(a, b)$ and $c + d$ in classical algebra. The neutral element for taking the maximum is $-\infty$, which is denoted by ε in max-plus algebra: $a \oplus \varepsilon = a,\ \forall a$. For addition, the neutral element is 0, in max-plus algebra denoted by e: $c \otimes e = c,\ \forall c$. Furthermore $\bigoplus_i a(i)$ denotes the maximum of all elements of $a(i)$ with respect to all appropriate indices i. Similarly, $\bigotimes_j c(j)$ is the sum of all elements of $c(j)$ with respect to all appropriate indices j.

The max-plus operations can also be extended to matrices. For the matrices $A, B \in \mathbb{R}^{m \times n}$ and $C \in \mathbb{R}^{n \times p}$ (with the matrix elements a_{ij}, b_{ij} and c_{jk} for $i = 1, ..., m$, $j = 1, ..., n$ and $k = 1, ..., p$) the following max-plus operations are defined:

$$(A \bigoplus B)_{ij} \equiv a_{ij} \oplus b_{ij}, \tag{3.22}$$

$$(A \bigotimes C)_{ik} \equiv (a_{i1} \otimes c_{1k}) \oplus ... \oplus (a_{in} \otimes c_{nk}) = \bigoplus_{j=1}^{n} (a_{ij} \otimes c_{jk}). \tag{3.23}$$

3.2.2 A Cyclic Railway Timetable in Max-Plus Algebra

The events in a railway system are the departures and arrivals. The departure of a train usually depends on one or more preceding arrivals and often also on another departure. For example, the departure of a train depends on the arrival time of the same train at the same station. When there are passenger transfers, the departure also depends on the arrival of the feeder trains. Furthermore, it has to wait for the preceding train that uses the same infrastructure. The timetable also provides an event: the planned departure time. Together this leads to a departure time that is equal to the *maximum* of the preceding event times *plus* the time needed for the process between the preceding event and the departure. This can be modeled as a system of max plus equations, as described by Goverde (2002).

For example, observe the departure of train 2024 from Utrecht (Ut) towards Gouda and The Hague. The train is not allowed to leave Utrecht before its planned departure time d_{Ut}^{2024}. Train 2024 already comes from Nijmegen, with a planned arrival time a_{Ut}^{2024} in Utrecht. There is a minimal dwell time of two minutes. Train 5624 coming from Zwolle acts as a feeder train for train 2024. Passengers need at least 4 minutes for this transfer. Furthermore, right after departure, train 2024 shares the same infrastructure with the earlier scheduled train 2824, running from Utrecht towards Gouda and Rotterdam. This means that train 2024 has to respect a two-minute headway after the departure of train 2824.

However, for the preceding activities the realization has to be observed and not the plan. The realized arrival and departure times are represented by $\tilde{a}$ and $\tilde{d}$ respectively.

Now, four restrictions for the realized departure time of train 2024 can be formulated in classical algebra:

$$\tilde{d}_{Ut}^{2024} \geq d_{Ut}^{2024}, \tag{3.24}$$

$$\tilde{d}_{Ut}^{2024} \geq \tilde{a}_{Ut}^{2024} + 2, \tag{3.25}$$

$$\tilde{d}_{Ut}^{2024} \geq \tilde{a}_{Ut}^{5624} + 4, \tag{3.26}$$

$$\tilde{d}_{Ut}^{2024} \geq \tilde{d}_{Ut}^{2824} + 2. \tag{3.27}$$

Now, equations (3.24) – (3.27) can be reformulated in max-plus algebra:

$$\tilde{d}_{Ut}^{2024} = d_{Ut}^{2024} \oplus (2 \otimes \tilde{a}_{Ut}^{2024}) \oplus (4 \otimes \tilde{a}_{Ut}^{5624}) \oplus (2 \otimes \tilde{d}_{Ut}^{2824}), \text{ or} \tag{3.28}$$

$$\tilde{d}_{Ut}^{2024} = \begin{pmatrix} e & 2 & 4 & 2 \end{pmatrix} \otimes \begin{pmatrix} d_{Ut}^{2024} \\ \tilde{a}_{Ut}^{2024} \\ \tilde{a}_{Ut}^{5624} \\ \tilde{d}_{Ut}^{2824} \end{pmatrix}, \text{but more natural} \tag{3.29}$$

$$\tilde{d}_{Ut}^{2024} = \begin{pmatrix} 2 & 4 & 2 \end{pmatrix} \otimes \begin{pmatrix} \tilde{a}_{Ut}^{2024} \\ \tilde{a}_{Ut}^{5624} \\ \tilde{d}_{Ut}^{2824} \end{pmatrix} \oplus d_{Ut}^{2024}. \tag{3.30}$$

Because the timetable is cyclic with a period of 60 minutes, these restrictions are similar for every hour. To be able to represent these restrictions in a general manner, not the actual train number, but the train line number and the cycle number are used. An additional distinction is made for the two directions of the line, with for example 2001 indicating all 2000-trains in the *odd direction* (from The Hague to

Arnhem/Nijmegen) and 2002 for the *even direction* (from Nijmegen/Arnhem to The Hague). In fact, the 2000-line runs twice per hour. Therefore additional numbers 2003 and 2004 could be used for the second half hour in the odd and even direction respectively.

The planned departure time of the 2024 from Utrecht is 9:03. Assuming that the train services start at 5:00 in the morning, 9:03 is just after the start of the 5th cycle and therewith $d_{Ut}^{2002}(5) = 9{:}03$, but more generally in terms of minutes past the start of services: $d_{Ut}^{2002}(k) = 3 + (k-1) \cdot 60$, which is 243 for the 2024. The planned arrival of the 2002 trains, which is exactly at the hour, can be written as $a_{Ut}^{2002}(k) = 0 + (k-1) \cdot 60$. Furthermore, the arrival of the 5602-line is at 3 minutes before the hour ($a_{Ut}^{5602}(k) = 57 + (k-1) \cdot 60$) and the planned departure of the 2802 is 1 minute past the hour: $d_{Ut}^{2802}(k) = 1 + (k-1) \cdot 60$.

Regarding the planned departure and arrival times within the hour, the following has to be observed: the departure from Utrecht of train line 2002 in period 5 depends on the arrival of the same train in period 5 and on the departure of train 2802 in period 5. However, it does not depend on the arrival of train 5602 in period 5, but on its arrival in period 4. Now equations (3.30) can be written in a cyclic format:

$$\tilde{d}_{Ut}^{2002}(k) = \begin{pmatrix} 2 & 4 & 2 \end{pmatrix} \otimes \begin{pmatrix} \tilde{a}_{Ut}^{2002}(k) \\ \tilde{a}_{Ut}^{5602}(k-1) \\ \tilde{d}_{Ut}^{2802}(k) \end{pmatrix} \oplus d_{Ut}^{2002}(k). \tag{3.31}$$

For an easier distinction between the cycles and an easier expansion of the model, the equations with different cycle offsets are separated. Note that the entry is filled with ε (equivalent for minus infinite) when the left-hand-side departure does not depend on the corresponding event.

$$\tilde{d}_{Ut}^{2002}(k) = \begin{pmatrix} \varepsilon & 4 & \varepsilon \end{pmatrix} \otimes \begin{pmatrix} \tilde{a}_{Ut}^{2002}(k-1) \\ \tilde{a}_{Ut}^{5602}(k-1) \\ \tilde{d}_{Ut}^{2802}(k-1) \end{pmatrix} \oplus \begin{pmatrix} 2 & \varepsilon & 2 \end{pmatrix} \otimes \begin{pmatrix} \tilde{a}_{Ut}^{2002}(k) \\ \tilde{a}_{Ut}^{5602}(k) \\ \tilde{d}_{Ut}^{2802}(k) \end{pmatrix} \oplus d_{Ut}^{2002}(k). \tag{3.32}$$

In fact, it is also possible that events depend on other events of 2 or even more periods earlier. In those cases, the number of terms in equation (3.32) can be increased until the maximum number of cycle offsets is reached. From hereon we assume that events only depend on events of at most 1 period earlier. Goverde and Soto y Koelemeijer (2000) also describe how the zero-order term can be eliminated from the equations.

The left hand side of equation (3.32) can be extended to a vector which includes multiple departures and arrivals, in particular all departures and arrivals throughout the network within one cycle. Let us call this vector with n elements $\tilde{x}(k)$. The two right-hand-side vectors containing the earlier events will also contain all the events for the appropriate cycle: $\tilde{x}(k-1)$ is introduced and $\tilde{x}(k)$ is repeated. Also the last term in the example, the planned departure time, will be extended and form a vector of planned arrival and departure times of size n: $x(k)$. Finally, the vectors containing the minimal time differences are replaced by matrices, which contain a column for each event time: $A(0)$ and $A(1)$, both of size $n \times n$. Element $A(0)_{ij}$ represents the minimal time that has to elapse between the event j in period k and event i in period k; $A(1)_{ij}$ represents the minimal time between event j in period $k-1$ and event i in period k. This leads to

$$\tilde{x}(k) = A(1) \otimes \tilde{x}(k-1) \oplus A(0) \otimes \tilde{x}(k) \oplus x(k). \tag{3.33}$$

Goverde and Soto y Koelemeijer (2000) show that this equation can be rewritten as a purely first order equation:

$$\tilde{x}(k) = \widehat{A(1)} \otimes \tilde{x}(k-1) \oplus x(k). \tag{3.34}$$

Now the *state matrix*, which relates two consecutive periods of the timetable, is defined as $A = \widehat{A(1)}$. The state matrix A follows from the dependencies in the timetable and the minimal process times. Although the train order, passenger transfers and rolling stock circulation are taken into account in the state matrix, this matrix is independent of the actual timetable.

3.2.3 Timetable Evaluation Using Max-Plus Algebra

Several reliability measures for timetables can be determined with max-plus algebra. These are described by Goverde (2002), Soto y Koelemeijer et al. (2000), Subiono (2000), and others. A short summary of the most common measures is given here. The software tool PETER (Goverde and Odijk, 2002) integrates these measures in a max-plus evaluation tool for timetables.

Minimal cycle time The minimal cycle time of a timetable is the time needed to complete the longest cyclic chain of processes in the periodic timetable. The minimal cycle time is equal to the eigenvalue λ of the state matrix A in terms of max-plus algebra:

$$A \otimes \underline{x} = \lambda \otimes \underline{x}. \tag{3.35}$$

The eigenvector $\underline{x}$ represents the corresponding timetable with cycle time λ. The timetable structure defined by the state matrix A can only be operated within a cycle time T when the minimal cycle time λ is less than T. When $\lambda < T$, the timetable corresponding to A is considered to be *stable*.

To determine the eigenvalue λ, several algorithms have been developed. Karp (1978) describes a polynomial algorithm, which only computes the eigenvalue λ. The *power-algorithm* (Braker and Olsder, 1993, Subiono et al., 1998) also calculates the eigenvector $\underline{x}$. Braker (1993) also reports on the *extended Karp algorithm*, which is both faster than the original algorithm of Karp, and faster than the power algorithm. However, this extended algorithm can not handle zero-order terms. Cochet-Terrasson et al. (1998) describe the *policy iteration algorithm*, which, although not proven, seems to outperform the other algorithms. Therefore the policy iteration algorithm was implemented in PETER.

Critical Circuit A critical circuit with respect to the minimal cycle time is a circuit through the timed event graph representing the timetable, for which the minimal cycle time λ is needed to complete the circuit. When there is a unique critical circuit, the minimal cycle time will decrease if one of the process times in the critical circuit is shortened or one of the dependencies in this circuit is canceled.

Throughput A measure derived from the minimal cycle time is the network throughput ρ. This measure indicates the percentage of the cycle time necessary to operate one cycle of the timetable.

$$\rho = \lambda / T. \tag{3.36}$$

The network throughput (applicable to a whole network) is comparable to the *capacity consumption* measure for a corridor described by the International Union of Railways (UIC, 2004).

Stability margin Another reliability measure is the stability margin Δ. This is the largest simultaneous increment of all processes for which the system is still stable, i.e. it can still be operated within the cycle time T.

The stability can be seen as the difference between the available time and the necessary time for all paths through the network. Note that many of these paths are longer than one timetable cycle. In classical algebra this can be written as

$$\max_{k=0,\ldots,M} (A(k) - k \times T), \tag{3.37}$$

where M is the maximum order of the A-matrix, i.e. the number timetable cycles elapsing for the longest path.

Now a max-plus definition has to be given for the polynomial matrix of a finite matrix series, $A(0)...A(M)$:

$$\mathbf{P}(A(y)) = \bigoplus_{k}(A(k) \otimes y^k). \tag{3.38}$$

Then we can rewrite (3.39) in max-plus algebra:

$$\max_{k=0,\dots,M}(A(k) - k \times T) := \bigoplus_{k=0}^{M}(A(k) \otimes T^{-k}) = \mathbf{P}(A(T^{-1})). \tag{3.39}$$

The stability margin $\Delta = -\mu$ can now be found by solving the eigenvalue problem given by

$$\mathbf{P}(A(T^{-1})) \otimes \underline{x} = \mu \otimes \underline{x}. \tag{3.40}$$

The critical circuit with respect to the stability margin is not necessarily the same as the critical circuit with respect to the minimal cycle time. This is because the stability margin also depends on the number of running trains in the circuit. The new critical circuit is the circuit with the least average buffer time Δ.

Recovery time The recovery time R_{ij} between events j and i, is the maximal delay of event j, which can be recovered before it has propagated to event i. In other words, it is the smallest amount of slack over all paths in the network from event j to event i. Or rephrasing it once more, it is the difference between the event times x_j and x_i minus the sum of the minimal process times for the longest path between the events j and i.

$$R_{ij} = x_j - x_i - \mathbf{P}(A(T^{-1}))_{ij}, \tag{3.41}$$

where M is the maximum order of the A-matrix, and $\mathbf{P}(A(T^{-1}))_{ij}$ is the longest path between the events j and i.

Delay Propagation The propagation of pre-determined delays can be analyzed using max-plus algebra. Inserting an initial delay vector in equation (3.34), the delay propagation can be determined recursively.

PETER The timetable evaluation tool PETER (Performance Evaluation of Timed Events in Railways) is based on max-plus algebra. This tool is described by Goverde and Odijk (2002), and is used by ProRail. PETER is able to analyze the critical circuit, to calculate the cycle time and throughput, to determine the stability margin, and to analyze recovery times and delay propagation. A link with the DONS database facilitates these analysis. Interaction with other timetabling tools and an interface including network graphs assist in the timetable analysis. Goverde and Odijk (2002) present an evaluation of the Dutch intercity network 2000-2001, consisting of 317 departure events.

Disadvantages of PETER Although the minimal cycle time and its critical circuit point at the timetable circuit with the smallest amount of slack, there is no direct relation with the stability of the network timetable. First of all, this critical circuit may consist of processes which are not or hardly disturbed. Other circuits with more slack, but larger disturbances, may perform worse. Secondly, the critical circuit may be the result of a peripheral train line which hardly influences other parts of the network.

The analysis of delay propagation is only possible for delays which are present at the start of the evaluation. No further disturbances are incurred over the evaluation period. The evaluated situations are very specific and do not represent real world situations.

Furthermore, PETER only evaluates a timetable with respect to the characteristics described above. No indications are given of how to improve this timetable, and an analysis with random disturbances is not possible at all.

3.3 Railway Simulation and SIMONE

Simulation is an evaluation method based on the repetitive imitation of a system, using a simplified model of that system.

Simulation is especially interesting for analyzing complicated and heavily interrelated processes, foremost in combination with stochastic disturbances. These are exactly the characteristics of busy railway networks.

Some advantages of the use of simulation for the quality assessment of railway systems are discussed in Section 3.3.2.

In this thesis SIMONE is used as simulation tool to evaluate the quality of a railway timetable. Therefore SIMONE is discussed in Section 3.3.3. Note, however, that SIMONE is one of many railway simulators (see Section 2.8). These different simulation tools usually have a lot in common, but can be different on many points as well.

3.3.1 Types of Simulation

Simulation models are applied to many different systems with many different characteristics. Therefore simulation models appear in many different forms. Law and Kelton (2000) describe three different classifications, which are briefly given here.

Static versus dynamic A static simulation model is a model in which time does not play a role. An example is Monte Carlo simulation (see also Law and Kelton (2000)). A dynamic simulation model shows how a system evolves over time.

Deterministic versus stochastic A deterministic simulation model does not contain any random components. In this case simulation is still very helpful, when the deterministic relations are too complex to be evaluated analytically. Many systems have some random influences. Simulation models including random components are called stochastic simulation models.

Continuous versus discrete In continuous simulation models, state variables change continuously over time, while in discrete simulation models state variables change instantaneously at a countable number of points in time. For discrete simulation models, also called discrete-event simulation models, two time advance mechanisms can be distinguished: *next-event time advance* and *fixed-increment time advance*. The names already give away the working of these mechanisms.

Furthermore, a distinction between synchronous and asynchronous simulation is found in the literature. In fact, it is used for two different classifications.
Synchronous versus asynchronous The terms synchronous and asynchronous simulation are first used to indicate the order of simulation (see e.g. Ghosh (1984)). In synchronous simulation models an event-list is maintained. Using this event-list, the model processes all events in chronological order. The time and memory costs of maintaining the event-list is saved in asynchronous simulation. In that case, events are not always processed in their natural order.

A more common use of the terms synchronous and asynchronous indicates the distinction between the simulation of planned timetables and randomly generated timetables (see e.g. Pachl (2002)). With synchronous simulation, predetermined timetables are evaluated. In asynchronous simulation, train paths are stochastically generated. Possible conflicts are solved by scheduling rules.

3.3.2 Advantages and Disadvantages of Simulation

Compared to other research methods of complicated stochastic systems, simulation has advantages and disadvantages. Some of the disadvantages are described below, but some of the advantages of simulation are discussed first.

- **Detailed modeling** Analytical models are often unable to include many details, because these details are incompatible with the model. Simulation models have the flexibility to incorporate virtually any level of detail. However, new details will increase the running time of the simulation.

- **Whole network** Analytical models often consider only one line or a restricted network; others have very restrictive assumptions about the model. Simulation, however, can deal with whole networks without these restrictive assumptions.

- **Complicated disturbance distribution** Most analytical models are based on very restrictive assumptions about primary disturbance distributions. Simulation offers the possibility to have any arbitrary distributions. It is also easy to use different distributions for different trains or locations.

- **Artificial situations** The main disadvantage of analyzing realization data is that only past situations can be considered. The reliability implications of new timetables or new infrastructure cannot be analyzed beforehand. The use of real life experiments is also impossible. Analytical models and simulations have more possibilities to look into the future.

- **Detailed output** The output of analytical models is often restricted to a few basic figures. The practical reasons and model characteristics behind the figures is often lost. Simulation models have the ability to create output in almost any desired format, giving information about the smallest modeled detail. However, the output of analytical models sometimes gives more information about relations in the railway system.

- **Animation** Animation is not only used because it looks fancy. It can also be very useful in detecting reasons for secondary delays. Furthermore, it makes both the model and the results easier to understand.

- **Computer speed** Computer speeds and capacities have improved so much over the past decades that it is possible to simulate much more and much faster than in the past. Consequently, the scientific possibilities have increased too.

Certainly, simulation has several disadvantages as well. A few are given here.

- **Only evaluation** The output of simulation models is just an evaluation of the simulated timetable. The results do not tell what a good timetable should look like, nor how to improve the involved timetable. In other words: it does not provide the user with analytical relations between parts of the system.
- **Very complex** The complexity of railway systems asks for high level programming for railway simulation tools. Moreover, the users of a simulation tool are required to have much knowledge about both the real life railway traffic and the simulation tool itself, to be able to use it adequately. Furthermore, often cumbersome preparations are needed before simulation runs can be carried out.
- **Input disturbances** Simulation models evaluate the timetable by calculating delay propagation throughout the network. This delay propagation is based on the disturbances which are defined by the user. However, the size and frequency of disturbances is usually not known. This means that simulation results are based on the assumptions made about the disturbance distributions, and not on real life disturbances.

3.3.3 The SIMONE Model

SIMONE (SImulation MOdel of NEtworks) is a simulation tool developed for ProRail, the Dutch railway infrastructure capacity planner, and NS, the main Dutch passenger railway operator. Since 1999, it is regularly used by these organizations in punctuality research. This includes both the comparison of different timetables as well as studies to assess the effects of new infrastructure. Moreover, it is used for the simulations which are discussed later on in this thesis. SIMONE is able to receive input from the timetabling tool DONS (see Section 3.1).

For a more formal and detailed description of all functionalities of the latest SIMONE version one is referred to the conceptual model (Bouwman et al., 2004)

Discrete-event simulation One of the main characteristics of SIMONE is that it is a discrete-event simulation model. When a train leaves one timetable point it is directed to the next one. There it appears as much later as the timetable indicates, adapted for possible disturbances or recoveries. This feature saves much time compared to continuous time simulation and offers the opportunity to simulate a whole railway network, like that of the Netherlands.

It also means that many complex details in railway systems can be taken into account. Interactions between trains, such as headway times on the tracks, platform occupations in the stations, and connections for travelers, are present in SIMONE. Also complicated disturbance distributions, depending on train type, infrastructure, time and other characteristics can be incorporated. Every desired infrastructure layout can be built, which allows one to research artificial situations as well.

Using the simulation classification given in Section 3.3.1, SIMONE can be characterized as a dynamic, stochastic, discrete and synchronous simulation model with next-event time advance.

The infrastructure A simulation-run with SIMONE starts with building the infrastructure, which is made up of timetable points and open tracks. This infrastructure layout is imported from DONS.

Timetable points exist in several forms, but most of them are either a station or a junction. Although these two have different functions in reality, they are identical for SIMONE.

Every timetable point has a set of *platform groups*. Each platform group has a capacity in length and in number of trains. Super platform groups can be present, to define the platform capacity more specifically. Every train that arrives at or passes a station is assigned to a platform group or super platform group in the timetable. At the moment of entering a platform group, it is checked whether capacity is available.

Open tracks are defined by their length, their number of tracks, and their capacity in number of trains. The latter needs some explanation: the block safety system, as it is used in the Netherlands and other European countries, has not been incorporated in SIMONE. Instead of space distances, time distances (or headways) and a limit on the number of trains per track are used.

Timetable points are situated at the end of one or more open tracks. Every parallel track of an open track has its own in-out point at the edge of a station, as shown in Figure 3.1. These in-out points are situated at a specific side of the platform groups (i.e. the *A*- or *B-side* of the station) and have a relative geographical location. This enables the software to detect possible conflicts between trains. For all in-out points minimal time-distances between trains are defined. Usually these headway times are defined network wide, but deviations can be defined for a specific station or track. Headway times can also be train type or even train line dependent. Note that these headway times at the in-out points are in fact also headway times for entering an open track.

The timetable The timetable for SIMONE is based on the cyclic timetable from DONS, and (in the sixty-minutes cycle situation) only specifies the minutes within the hour. For each train line, for both directions, the timetable specifies the arrival, departure and passing times at all timetable points. The timetable defines, besides

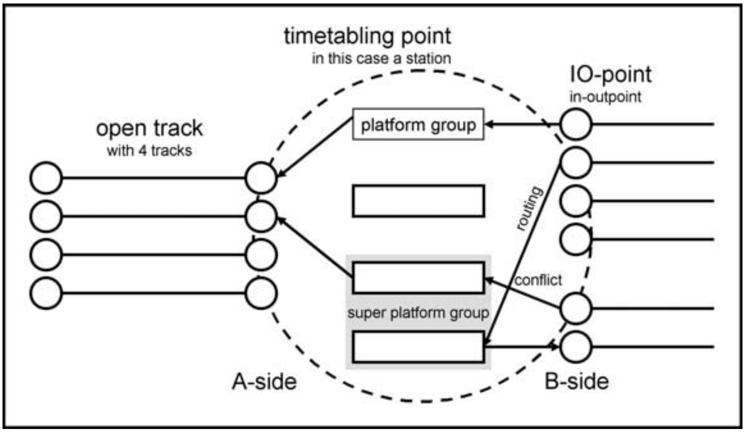


Figure 3.1: Layout of a timetable point connected to three open tracks.

these times, for each arrival the arrival in-out point and arrival platform group. The platform group and the departure in-out point are defined for each departure. Furthermore the timetabled time for the next process (a dwell time for arrivals or a running time for departures), and the accompanying dwell time supplement or running time supplement are given. For passing times, the incoming and outgoing in-out points are defined, as well as the upcoming running time and running time supplement.

Additional features The SIMONE simulation incorporates more than only the infrastructure and the timetable. Much operational information is defined on network level, timetable point level or train level.

- **Train length** The physical train length for each train line is defined to determine the occupation of platforms.
- **Passenger connections** For passenger connections it can be defined that certain departing trains have to wait for certain delayed arriving trains. This is to enable passengers to transfer. A maximum waiting time is also included.
- **Rolling stock layovers** For the rolling stock of each train line, a sequential train line can be defined. In the case that there is such a layover, a minimal layover time has to be respected.
- **First and last dwelling time** For trains without a layover at the start, the first dwelling time determines at which time the train *arrives* at the starting station and occupies (part of) its platform group. Likewise, the last dwelling

time determines the time at which time the arrival platform group at the final station is left.

- **Conflicts within timetable points** In SIMONE trains do not have an exact route through a station, but are assigned to an arrival in-out point, a platform group and a departure in-out point. Therefore, stations are referred to as *grey boxes*. Conflicts within the station area are imported from DONS. There they are deduced as shown in Figure 3.1: for each train a routing is drawn from the in-out point to the appropriate platform group and on to the in-out point. When these lines cross each other, a conflict will emerge, when both movements take place at the same moment.

Disturbances In SIMONE, the disturbances have only one cause: the user definition of the simulation run. A disturbance can be either a *dwell time disturbance*, a *departure disturbance*, or a *running time disturbance*. For all disturbances, a probability of occurrence and a distribution are defined. Furthermore a choice has to be made between *absolute* and *relative* disturbances. Without further specifications, the disturbances apply to all dwellings, departures, or running times, but they can also be more specific.

Every time a train enters an open track, there is a defined probability for a running time disturbance. If there is a disturbance, the size is randomly picked from the specified distribution. If an absolute disturbance is chosen, the running time is increased by this random disturbance. In the case of relative disturbances, the disturbance-percentage is multiplied with the minimal running time for the track where the train is disturbed. A disturbance on a certain trip can be compensated by the running time supplement on the same trip.

Dwell time and departure disturbances are almost always absolute disturbances. Like running time disturbances, there is a probability of occurrence and a probability distribution for the size of the disturbance. The difference between a dwell time disturbance and a departure disturbance is that a dwell time disturbance extends the dwell time, while a departure disturbance only effects the departure: the latter is not incurred until the train is ready to leave the station.

Secondary delays Secondary delays do not occur in an undisturbed simulation experiment; at least not when the timetable is conflict free. Secondary delays only appear because other trains are delayed already. Both primary delayed trains and secondary delayed trains can cause additional secondary delays.

Secondary delays occur when:

- there is no capacity in the assigned platform group, either in the number of trains or in the total train length;
- there is a conflict with another train between the platform and the in-out point (or vice versa);
- the headway of the previous train at the beginning of the track has not yet elapsed (on entering the track);
- the maximum number of trains on a track was already reached (on entering the track);
- the headway of the previous train at the end of the track has not yet elapsed (on leaving the track);
- the feeder train for a passenger connection has not yet arrived;
- the preceding train in the rolling stock circulation has not yet arrived.

Use of supplements Time supplements can be available both in the dwell times and in the running times. Running time supplements at a certain track can already be used to compensate for delays incurred at the same track. Likewise, dwell time supplements at a certain station can be used to make up for dwell time disturbances incurred at the same station, but they cannot be used to make up for departure disturbances. Supplements will only be used when a train is delayed. Trains will therefore never arrive or depart early.

When the planned running (or dwell) time is smaller than the minimal running (or dwell) time, we speak about negative supplements. Negative supplements create disturbances, which have the (absolute) size of the supplement. Negative supplements do occur sometimes because of rounding to full minutes. When timetables are created in an acceptable manner, these negative supplements are small, and are compensated by larger positive supplements not much later.

Dispatching and limited disturbances Unless dispatching rules have been implemented, which has to be done per timetable point, SIMONE follows the first-come-first-serve principle.

Only a very limited traffic control function has been included in SIMONE. This has two reasons. Traffic control is most important when disturbances of the railway traffic are large. However, no traffic scenario is so robust that it can absorb very large disturbances. Therefore it is less interesting to compare scenarios for such circumstances. Besides that, it is very hard to define general traffic control rules, because these are very situation specific. On top of that, different dispatchers may act differently in comparable situations. This also implies that all research done with SIMONE is based on experiments with relatively small disturbances.

Animation Another important feature of simulation is animation. However, it should only be used for visual verification and presentation, because it slows down the simulations considerably. Animation in SIMONE is available on two levels. Figures 3.2 and 3.3 show the Netherlands and greater Amsterdam on a network level. On the network level it is also possible to show updated punctuality figures on the map.

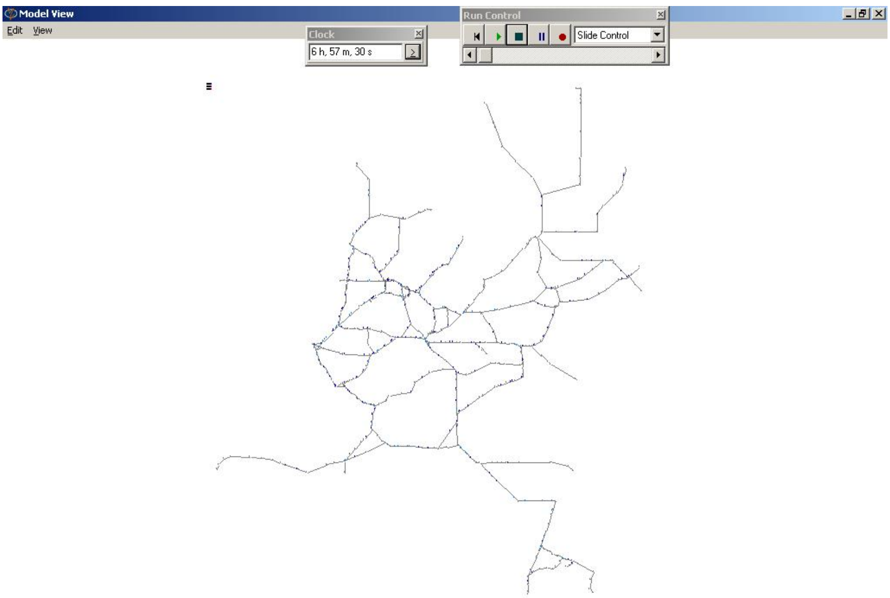


Figure 3.2: A screenshot of a SIMONE animation of the whole Dutch network.

On the lower level one can view the more detailed animation of the timetable points and open tracks. For a timetable point one can choose between animation of the platform group consumption and animation of conflicting routes in the timetable point area.

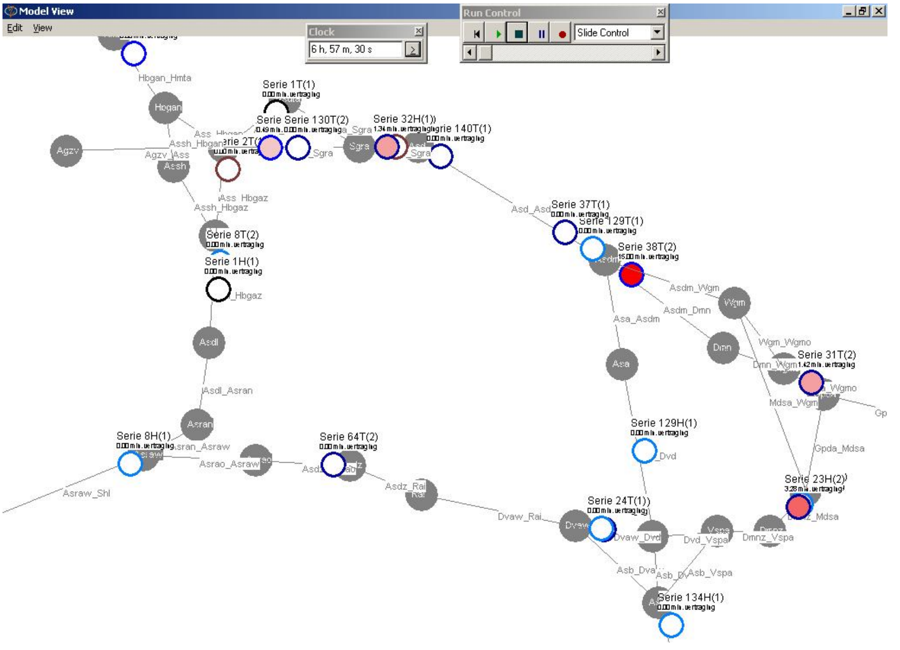


Figure 3.3: A screenshot of a SIMONE animation of the Amsterdam area.

Some disadvantages Although many features are available in SIMONE, not everything can be modeled. Some characteristics which have not been modeled, but can have their influence on the punctuality are given here.

- **Stations are not modeled in detail** The exact infrastructure layout of stations has not been modeled in SIMONE. This means that not all relations and dependencies between train routes and platform occupations have been modeled in full detail. This implies that delay propagation can be under- or overestimated. The new simulation tool FRISO, described in the next section, is developed to analyze stations in more detail.

- **No personnel schedules** Simulations in SIMONE are based on one-hour timetables, which is in line with the real timetable in the Netherlands. The circulation of rolling stock is taken into account in SIMONE. This is straightforward, because rolling stock schedules are mostly cyclic as well. However, personnel schedules do not show these cyclic patterns. This makes it most difficult to implement.

- **No dispatching** The first-come-first-serve principle has been implemented in almost all circumstances. This means that no dispatching rules have been implemented. However, this choice was made, because SIMONE has been developed for the analysis of small delays, and dispatching only plays an important role when delays are larger. Besides the passenger connections which can be modeled in SIMONE, some more waiting and dispatching rules are implemented in the DVM-mode of SIMONE, which is described in the next section.

3.3.4 SIMONE-DVM and FRISO

The SIMONE-simulation software was developed to tackle a wide range of problems. Besides scenario comparisons, it can also be used for scientific research of timetable characteristics. Still, one simulation tool is never able to answer all questions satisfactorily. Two specific wishes of SIMONE-users have led to a functional extension of the original SIMONE, SIMONE-DVM, and the development of the more detailed simulation tool FRISO. Although they are both not applied in this thesis, they are shortly described here.

SIMONE-DVM Dynamic Traffic Management (Dynamisch VerkeersManagement or DVM) is an innovative planning principle to improve the reliability of highly utilized railway networks such as in the Netherlands. In DVM, trains are not scheduled as a straight line in the time-distance diagrams with exact times for departures and arrivals, but as *time windows*. The lower bound of this window indicates the earliest departure and arrival times, the upper bound the latest departure and arrival times. The published timetable should state the earliest departure times and latest arrival times. A clear distinction between the operational plan and the published timetable arises here.

It is even possible to schedule overlapping time windows, which means that timetables are not necessarily conflict free anymore. Schaafsma (2001) describes the DVM-ideas and its technical details more precisely. A DVM-mode has now been incorporated in the standard SIMONE software. SIMONE-DVM can therefore classified identically as SIMONE.

FRISO For more detailed simulations of large railway nodes, a detailed version of SIMONE, FRISO, is being developed. FRISO stands for Flexibele Rail Infra Simulatie Omgeving (in English: Flexible Rail Infrastructure Simulation Environment). The main difference between SIMONE and FRISO is the level of detail of the infrastructure. The regular SIMONE only has tracks and stations (or nodes), and only models time distances between trains. FRISO incorporates the block occupation safety system, and monitors train speeds and acceleration rates continuously. This means that every switch and every signal is present in the model. Due to the level of detail and the accompanying increased simulation time, FRISO is only applicable to a restricted network. A precise description of the working of FRISO is given by Loeve et al. (2005).

As opposed to SIMONE-DVM, which is integrated in SIMONE, the detailed simulations with FRISO are executed with the separate FRISO software. Still, the availability of FRISO increases the possibilities of researchers. Foremost it enables them to look in more detail at, for example, bottlenecks in the network.

FRISO can be classified as a dynamic, stochastic, continuous and synchronous simulation model with next-event time advance.

Chapter 4

Running Time Supplements

To attain an acceptable level of punctuality of the train services, technically minimal running times are increased with running time supplements in the published timetables. These supplements are used to decrease, or even eliminate, incurred delays. For the operation of the timetable, it is both important to have sufficient running time supplements, and to have the supplements at the right location and at the right moment.

In the first section of this chapter, general ideas about running time supplements are discussed. An analytical approach to optimize the supplement allocation is presented in Section 4.2. For larger problems, a numerical allocation model is described in Section 4.3. Then a practical case for the Haarlem–Maastricht corridor is worked out in Section 4.4. Some concluding remarks and a discussion on practical implications can be found in Section 4.5.

4.1 General Ideas behind Running Time Supplements

4.1.1 Total Size of the Supplements

To obtain a high reliability of train services, it is desirable to be able to run faster than scheduled, to make up for earlier delays. This means that scheduled running times should be longer than the technically minimal running times. The (positive) difference between the scheduled running time and the technically minimal running time is called *running time supplement*.

The International Union of Railways (UIC, 2000) has published Leaflet 451-1 on the size of running time supplements. In their recommendations, the supplements are the sum of a distance dependent supplement and a percentage of the technically minimal running time. The distance dependent supplement is 1.5 min/100 km for locomotive-hauled passenger trains and 1 min/100 km for multiple unit passenger trains. The running time dependent supplements vary between 3% for relatively slow trains and 7% for faster trains. For locomotive-hauled trains, this percentage also depends on the total weight of the train. Supplements for cargo trains are generally higher. Furthermore, the running time dependent supplement can be replaced by a second distance dependent supplement in some cases.

In the Netherlands, running time supplements are approximately 7% of the technically minimal running times. This percentage is used nationwide, for all types of passenger services. However, due to roundings (because of the integer timetable) and local circumstances, the actual percentage may deviate slightly. Furthermore, cargo trains are planned differently. First of all, there is a difference between the (lower) planning speed and (higher) possible speed of cargo trains. On top of that, the planned acceleration and deceleration are based on conservative numbers of the locomotive power and the train weight. Usually trains are less heavy than planned, and can therefore accelerate and decelerate faster. In any case, the supplements are, more or less, proportionally allocated with respect to the minimal running time.

In Switzerland, running time supplements have several components (see Haldeman, 2003). First of all, there is a relative running time supplement, which is 7% of the running time for passenger trains and 11% for cargo trains. Secondly, Special Operational Supplements are added at, for example, highly utilized nodes. Additionally, one minute of supplement is added for each 30 minutes of running time. For trips with high average speeds the supplements have to be larger.

In the United Kingdom, planned running times are based on past performance on the particular railway section (Rudolph, 2003). Supplements are not explicitly defined, but included in the running times.

In general, higher running time supplements lead to a better reliability of train services. However, higher supplements also lead to higher planned running times. This means for the passengers that travel times increase. Furthermore, given the present day block system in the Netherlands, which is comparable with most safety systems in Europe, longer running times increase the block-occupation time and therewith the track or station consumption. Additionally, longer running times require more personnel and rolling stock.

4.1.2 Allocation of Supplements

In this section different ideas about the allocation of running time supplements are discussed. It is assumed that the total size of the supplements over the network is always the same. This means that larger supplements at one place always have to be compensated by smaller supplements at another place. Always remember that trains with decreased supplements will experience more delays and can cause many secondary delays, also to trains that have increased supplements.

Additionally, the measuring locations of punctuality and average delay can have their influence on the supplement allocation. For passenger satisfaction, one intuitively has to weigh the arrival delays with respect to the number of arriving passengers. Still, official performance measures are usually based on the arrivals at certain large stations only, and they are not weighted for the number of passengers. Timetable optimization with respect to the passengers is in this case quite different from optimization with respect to the official performance measure.

Below some general considerations are given for the allocation of running time supplements, both between lines and within one line.

- **More supplements for trains with more passengers** It is obvious that more passengers will arrive punctually when trains with most passengers are given more running time supplements. But these same passengers are supplied with higher planned travel times.
- **More supplements just before nodes** In this situation all supplements between two nodes are shifted to the last part of this track. This means that the first supplements are not lost when no disturbances occur in this part. They can still be used for disturbances later on. When disregarding secondary delays, the expected delay will decrease at the nodes. This has the advantage that the probability of causing secondary delays is smaller at the nodes. A consequence is that, when disturbances occur earlier on the track, trains cannot catch up with their schedule until just before the nodes. This implies that more lateness is expected at intermediate, usually smaller, stations. This shift in supplements can even be extended to the whole train line. In this case, it can be argued that the smallest possible delays are achieved at the endpoint of the line, if secondary delays are not taken into account. This also means that sequential services, performed by the same rolling stock, have the smallest expected departure delay. However, when this principle of collecting supplements is used on tracks which include junctions or important nodes, there is an increased probability for secondary delays due to the increased intermediate lateness. So shifting supplements past junctions or nodes can have negative implications. Shifting supplements past measuring points has a comparable result: the intermediate lateness, at the measuring points, increases.

- **More supplements for long lines** It is often argued that long lines need more running time supplements. One reason is that long lines have a higher probability of accumulating delays. Additionally, incurred delays are propagated over longer distances. This means that more other trains can be disturbed as well.

- **More supplements for intercity trains** Intercity and high speed services are the backbone of a railway system. Therefore the quality of these services is considered more important than that of other services. As punctuality is one of the most important quality measures, increased supplements for these trains can help to achieve this goal. On the other hand, the increased supplements lead to longer travel times, not only decreasing the quality of the train itself, but also decreasing the speed or quality difference with other services. Still, these decreased running time differences have an additional positive influence on punctuality. This homogenizing effect can be used to increase headways between trains, which decreases the effect of secondary delays, as described in Chapter 5. Finally it should be mentioned that intercity lines are often the longer lines and usually carry more passengers. Therefore the deliberations here overlap with those of *long lines* and *more passengers*.

- **More supplements on intensively utilized tracks** Where the capacity consumption of the network is high, there is an increasing risk for secondary delays. Not only the headways are smaller, which increases the probability of knock-on delays, but also the number of affected trains is higher in case of disturbances. Therefore, a high punctuality is more important on these tracks. This reasoning can also be applied to intensively utilized stations in the network.

- **More supplements during rush hours** Longer boarding and alighting times cause additional delays during rush hours. Due to the higher weight of the longer and more crowded trains, acceleration and breaking also takes more time. Therefore longer planned running times during rush hours are preferred. Furthermore, during rush hours there are also more trains. This means that more secondary delays are expected.

- **More supplements where disturbances are larger** The analysis of historical data of railway traffic can provide information on where disturbances occur. To minimize propagation of delays, it would be best to absorb these disturbances immediately with supplements. Therefore, to decrease total delays, it is most useful to match the supplements with the disturbances. However, historical data usually only provides information on the actual delays, not separated in primary and secondary delays.

4.2 Analytically Minimizing Average Delay

In this section we consider one train, from 0 to N, which travels along N sequential trips, as in Figure 4.1: trip 1 from station 0 to station 1, trip 2 from station 1 to station 2, and so on until trip N from station $N-1$ to station N. Each trip has its own disturbance distribution and the total amount of running time supplements is given. It is assumed that these distributions are known. The objective is to allocate the supplements in such a way, that the average delay is minimal. At the intermediate stations 1 to $N-1$ there are neither disturbances, nor dwell time supplements. This implies that the departure delay from a certain station is equal to the arrival delay at the same station. We assume that departures from the first station are on-time.

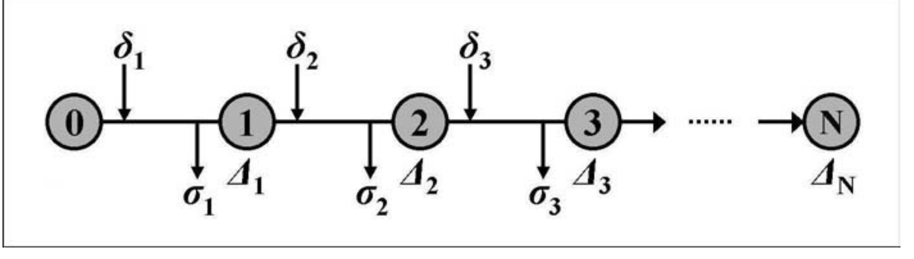


Figure 4.1: A train line from station 0 to station N is regarded, where trains incur a disturbance δ_i on the running time from station $i-1$ to station i. The running time supplement σ_i can be used to recover (part of) this disturbance.

The trips can be viewed in several ways. First of all, one can consider the intermediate points as minor stations on a short train line. Secondly the problem can be considered as a long train service for which only the large stations are considered and the other intermediate stations are skipped for evaluation. Depending on how the problem is viewed, the average delay can be weighted for the different intermediate stations. Note that the following important assumption is made: disturbances incurred on a certain trip can already be reduced by the running time supplement on that same trip.

4.2.1 Results for Exponentially Disturbed Trains on Two Trips

Consider a train consisting of two trips: from station 0 via station 1 to station 2, as shown in the left part of Figure 4.1. The train incurs a disturbance from an exponential distribution with an average size of $\frac{1}{\lambda_1}$ on the first trip and $\frac{1}{\lambda_2}$ on the second trip. This choice for an exponential distribution facilitates further analysis, and leads to the disturbance density functions

$$\begin{aligned} f_1(\delta) &= \lambda_1 e^{-\lambda_1 \delta}, \quad \text{for } \delta \geq 0, \quad 0 \text{ otherwise}, \\ f_2(\delta) &= \lambda_2 e^{-\lambda_2 \delta}, \quad \text{for } \delta \geq 0, \quad 0 \text{ otherwise}, \end{aligned} \tag{4.1}$$

for the first and second trip respectively. The total running time supplement σ_T has to be divided between the two trips in such manner that the average arrival delay at stations 1 and 2 is minimized. Define σ_1 and σ_2 as the supplements for the two trips. Then σ_1 and σ_2 should be non-negative, and, of course, $\sigma_2 = \sigma_T - \sigma_1$.

First, the probability that the arrival delay Δ_1 at station 1 is less than Δ is derived. This is the case when the disturbance δ_1 is smaller than Δ plus the supplement σ_1 on this trip:

$$P(\Delta_1 \leq \Delta) = \int_0^{\sigma_1 + \Delta} \lambda_1 e^{-\lambda_1 x} \mathrm{d}x = 1 - e^{-\lambda_1(\sigma_1 + \Delta)}, \text{ for } \Delta \geq 0, \quad 0 \text{ otherwise.} \tag{4.2}$$

Differentiation leads to the density function $h_1(\Delta) = \lambda_1 e^{-\lambda_1(\sigma_1+\Delta)}$ for the arrival delay at station 1, and an average arrival delay at station 1 of

$$E\Delta_1 = \int_0^{\infty} x \cdot \lambda_1 e^{-\lambda_1(\sigma_1 + x)} \mathrm{d}x = \frac{1}{\lambda_1} e^{-\lambda_1 \sigma_1}. \tag{4.3}$$

For evaluation of the second trip, both disturbance distributions and running time supplements have to be taken into account. Figure 4.2 shows the integration bounds which have to be used to find the delay density at the second station: a disturbance on the first trip can be recovered by the supplements on both trips, but a disturbance on the second trip can only be recovered by a supplement on the second trip.

This leads to the following equation for the probability that the arrival delay at station 2 is less than or equal to Δ:

$$\begin{aligned} P(\Delta_2 \leq \Delta) &= \int_0^{\sigma_T - \sigma_1 + \Delta} \lambda_2 e^{-\lambda_2 y} \int_0^{\sigma_T + \Delta - y} \lambda_1 e^{-\lambda_1 x} \, \mathrm{d}x \mathrm{d}y \\ &= 1 - e^{-\lambda_2(\sigma_T - \sigma_1 + \Delta)} + \frac{\lambda_2}{\lambda_1 - \lambda_2} e^{-\lambda_1(\sigma_T + \Delta)} \left[1 - e^{(\lambda_1 - \lambda_2)(\sigma_T - \sigma_1 + \Delta)} \right] \end{aligned} \tag{4.4}$$

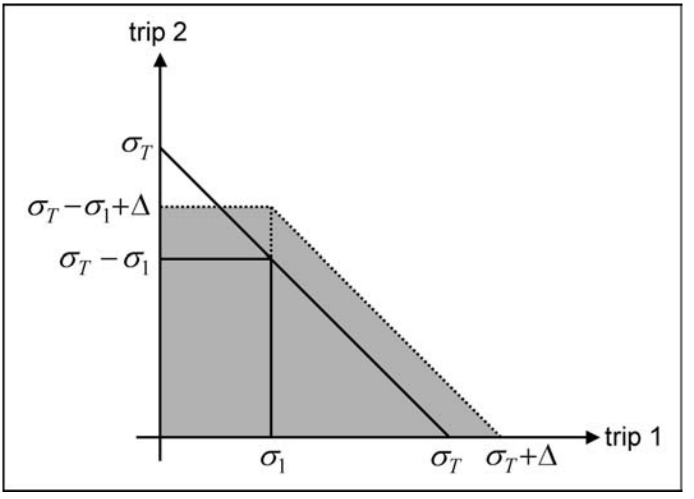


Figure 4.2: The integration upper bounds for the second trip.

Differentiation with respect to Δ leads to the density function of the arrival delay at station 2:

$$\begin{aligned} h_2(\Delta) = {} & \lambda_2 e^{-\lambda_2(\sigma_T-\sigma_1+\Delta)} + \frac{\lambda_2}{\lambda_1-\lambda_2}[-e^{-\lambda_1(\sigma_T+\Delta)}(\lambda_1-\lambda_2)e^{(\lambda_1-\lambda_2)(\sigma_T-\sigma_1+\Delta)} - \\ & \lambda_1 e^{-\lambda_1(\sigma_T+\Delta)}(1-e^{(\lambda_1-\lambda_2)(\sigma_T-\sigma_1+\Delta)})] \\ = {} & \lambda_2 e^{-\lambda_2(\sigma_T-\sigma_1)} e^{-\lambda_2\Delta}(1-e^{-\lambda_1\sigma_1}) - \\ & \frac{\lambda_1\lambda_2}{\lambda_1-\lambda_2}[e^{-\lambda_1\sigma_T}e^{-\lambda_1\Delta} - e^{-\lambda_2(\sigma_T-\sigma_1)}e^{-\lambda_1\sigma_1}e^{-\lambda_2\Delta}]. \end{aligned}$$

Then $E\Delta_2 = \int_0^\infty \Delta \cdot h_2(\Delta)\,\mathrm{d}\Delta$ or

$$\begin{aligned} E\Delta_2 = {} & \frac{1}{\lambda_2}(1-e^{-\lambda_2\sigma_1})e^{-\lambda_2(\sigma_T-\sigma_1)} - \\ & \frac{\lambda_2}{\lambda_1(\lambda_1-\lambda_2)}e^{-\lambda_1\sigma_T} + \frac{\lambda_1}{\lambda_1(\lambda_1-\lambda_2)}e^{-\lambda_2(\sigma_T-\sigma_1)-\lambda_1\sigma_1}. \end{aligned}$$

To minimize the unweighed expected average arrival delay at stations 1 and 2, we have to minimize $E\overline{\Delta}$ or $\frac{1}{2}(E\Delta_1+E\Delta_2)$. Differentiation and solving the resulting equation for σ_1 only leads to an implicit equation for the optimal value of the supplement on the first trip σ_1^*:

$$e^{(\lambda_1+\lambda_2)\sigma_1} - e^{\lambda_2\sigma_1} = e^{\lambda_2\sigma_T}. \tag{4.5}$$

The value of σ_1^* cannot be derived from this equation explicitly, but it can be approximated iteratively. To do this, rewrite equation (4.5) as $X^p = Q+X$ or $X = \sqrt[p]{Q+X}$, with $Q \equiv e^{\lambda_2 \sigma_T}$, $X \equiv e^{\lambda_2 \sigma_1}$, and $p \equiv 1 + \frac{\lambda_1}{\lambda_2}$.
If we then take $k_0 = \sqrt[p]{Q}$ and $k_{i+1} = \sqrt[p]{Q + k_i}$ we obtain an increasing and bounded sequence, with $\lim_{i\to\infty} k_i = X$, so $\sigma_1^* = (\ln X)/\lambda_2$. This approximation algorithm has a fast convergency to X.

A lower bound for σ_1^* can also be calculated. Equation (4.5) tells that $e^{(\lambda_2+\lambda_1)\sigma_1} > e^{\lambda_2 \sigma_T}$, which means that $(\lambda_1+\lambda_2)\sigma_1 > \lambda_2 \sigma_T$. This gives us a lower bound on σ_1^* and an upper bound on σ_2^*:

$$\sigma_1^* > \frac{\lambda_2}{\lambda_1+\lambda_2}\sigma_T, \text{ and } \sigma_2^* < \frac{\lambda_1}{\lambda_1+\lambda_2}\sigma_T.$$

Optimizing Punctuality Instead of minimizing the average delay, it is also possible to maximize the average punctuality. A train is considered punctual when its arrival is at most D minutes delayed. Therefore we try to maximize $\frac{1}{2}[P(\Delta_1 \leq D) + P(\Delta_2 \leq D)]$. These two cumulative distribution functions are already given in equations (4.2) and (4.4). To maximize the sum of these equations, we take the derivative and solve the respective equation with respect to σ_1. This leads to almost the same equation as for the delay minimization objective. There is only a new factor included in the right hand side:

$$e^{(\lambda_1+\lambda_2)\sigma_1} - e^{\lambda_2\sigma_1} = \frac{\lambda_1}{\lambda_2} e^{(\lambda_2-\lambda_1)D} \cdot e^{\lambda_2\sigma_T}. \tag{4.6}$$

The optimal value for the supplement on the first trip can be found with the same approximation algorithm as before, but now with $Q \equiv \frac{\lambda_1}{\lambda_2} e^{(\lambda_2-\lambda_1)D} \cdot e^{\lambda_2\sigma_T}$.

Moreover, when $\lambda_1 = \lambda_2$ equations (4.5) and (4.6) are identical. Even more surprisingly, the maximization of the punctuality is independent of the punctuality margin D when the exponential distributions are equal.

Two trips with equal disturbances Next consider the same problem again, but now with the same exponential disturbance distribution with average $\frac{1}{\lambda}$ on both trips. This implies $\lambda_1 \equiv \lambda_2 \equiv \lambda$. When using the equations above, one gets $E\Delta_1 = \frac{1}{\lambda}e^{-\lambda\sigma_1}$ and a division by zero for $E\Delta_2$. However, an explicit function for $E\Delta_2$ can be found by applying De l'Hôpital's rule (Apostol, 1967):

$$E\Delta_2 = \frac{1}{\lambda} e^{-\lambda\sigma_T}[e^{\lambda\sigma_1} + \lambda(\sigma_T - \sigma_1) + 1].$$

To find the optimal σ_1^*, equation (4.5) can be used again to find $(e^{\lambda\sigma_1})^2 - e^{\lambda\sigma_1} - e^{\lambda\sigma_T} = 0$, and the abc-formula can be applied to solve for $e^{\lambda\sigma_1}$.

The optimal supplement for the first trip is now found to be

$$\sigma_1^* = [\ln(1 + \sqrt{1 + 4e^{\lambda\sigma_T}}) - \ln 2]/\lambda,$$

and for the second trip $\sigma_2^* = \sigma_T - \sigma_1^*$. From this, the following insights can be deduced:

- When σ_T is less than $\frac{1}{\lambda}\ln 2$, $\sigma_1^* > \sigma_T$, and $\sigma_2^* < 0$. To avoid negative supplements, $\sigma_1^* = \sigma_T$ and $\sigma_2^* = 0$, when $\sigma_T < \frac{1}{\lambda}\ln 2$.
- For any $\sigma_T > 0$, $\sigma_1^* > \sigma_2^*$
- For $\sigma_T \to \infty, \sigma_1^* \downarrow \frac{1}{2}\sigma_T$ and $\sigma_2^* \uparrow \frac{1}{2}\sigma_T$.

General results Furthermore, the following results can be found easily with these equations. Assume that $\lambda = 1$ and the total supplement is equal to the total average disturbances, so $\sigma_T = 2$. Then the optimal supplements are $\sigma_1^* \approx 1.183$ and $\sigma_2^* \approx 0.817$. This leads to $E\Delta_1 \approx 0.306$, $E\Delta_2 \approx 0.688$, and an overall average expected delay of 0.497. When we compare this with the situation where $\sigma_1^* = \sigma_2^* = 1$, a slight reduction of the average delay is found: -1.2%. When σ_T is doubled to 4, $\sigma_1^* \approx 2.068$. This already indicates a fast convergence of σ_1^* to 50% of σ_T: it was 59.1% for $\sigma_T = 2$ and already decreased to 51.7% for $\sigma_T = 4$. On the other hand, $\sigma_1^* \approx 0.799$ (or 79.9%) for $\sigma_T = 1$, and as remarked earlier, 100% of σ_T, for $\sigma_T \leq \ln 2$.

The solid line in Figure 4.3 shows the optimal supplement size for the first trip, σ_1^*, as a function of the total supplement σ_T, given an average disturbance of 1 on both trips. Additionally the broken line shows σ_1^* as fraction of σ_T. For a small value of σ_T, σ_1^* is 100% of σ_T, but this percentage decreases when σ_T exceeds $\ln 2$, and when σ_T goes to infinity, σ_1^* drops to 50% of σ_T. The asymptote $\sigma_1^* = 0.5 \cdot \sigma_T$ is also shown.

In Section 4.3, a numerical model is introduced to optimize the supplement allocation for more than 2 trips. It is shown that the delay decreases are much larger in the case of more trips.

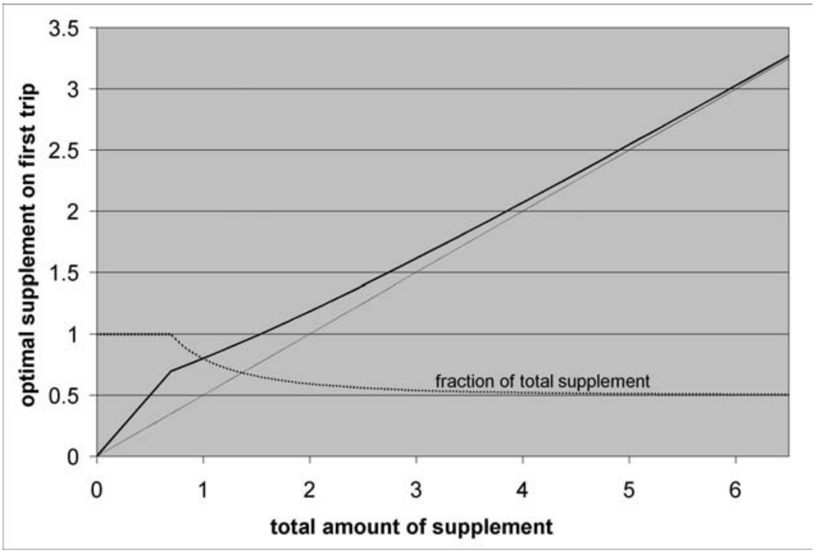


Figure 4.3: σ_1^* is the optimal size of the supplement for a given size of total supplement σ_T and $\lambda = 1$.

4.3 Numerical Optimization

For a general number of trips, we can calculate the optimal supplement allocation by discretization of the disturbances. When the disturbances are discretized, and a maximum delay is introduced, probability vectors can be calculated. The discretization steps should be small enough for the desired precision. The maximum delay has to be so large that the probability of larger delays is negligible. For each trip there is a disturbance vector. This vector contains the probabilities for the disturbance to have a certain size. Taking the convolution of the disturbance vector with the arrival delay of the previous trip leads to the total delay vector. After correcting for the supplement on the trip, the arrival delay vector at the end of the trip is found. Recursive calculations give the arrival delay distributions at all stations.

The accuracy of the result mainly depends on two parameters: the maximum delay Δ^{max} which is taken into account, and the size of the discretization-interval, V. These two parameters together give the total number of intervals: $I = \Delta^{max}/V$, excluding the on-time category. Δ^{max} should be chosen such that the probability of delays larger than Δ^{max} is 0 or negligible.

For the optimization of the supplement allocation over these trips, 2-OPT is used. For the initial allocation, the average supplement or any better guess for the allocation can be chosen. The approximation described here is $O(I^2)$.

Determining the expected average delay for a given supplement allocation
First, we must be able to calculate the average expected delay for any supplement allocation. For now assume that the supplements per trip are known and equal $\sigma_1, \sigma_2, ..., \sigma_N$, with N the number of trips.

Secondly the *disturbance density functions* $f_j(i)$ for the different trips ($j = 1, ..., N$) are discretized as follows:

$$\begin{aligned} f_j(0) &= P(\delta_j \leq \tfrac{1}{2}V), \\ f_j(i) &= P((i-\tfrac{1}{2})V \leq \delta_j \leq (i+\tfrac{1}{2})V), \text{ for } i = 1, ..., I, \end{aligned} \tag{4.7}$$

with δ_j the disturbance on trip j.

Now the density functions of the arrival delay $h_j(i)$ at station j have to be calculated for $j = 1, ..., N$. To do so, the density functions $g_j(i)$ are introduced, which represent the delay densities before the reduction by the supplement σ_j.

The arrival delays can be calculated in an iterative process, where each $g_j(i)$ is based on the arrival delay probability density $h_{j-1}(i)$ of the preceding measuring point. Indeed,

$$g_j(i) = \sum_{k=0}^{i} h_{j-1}(k) \cdot f_j(i-k). \tag{4.8}$$

Note that all departures from station 0 are on-time under the model assumptions, leading to $h_0(0) = 1$, and $h_0(i) = 0$ for $i = 1, ..., I$. However, any discrete departure distribution can be assigned to the vector $h_0(i)$.

Next, the supplement is deducted from the density $g_j(i)$, to obtain the arrival density $h_j(i)$ at station j:

$$\begin{aligned} h_j(0) &= \sum_{k=0}^{\sigma_j/V} g_j(k), \\ h_j(i) &= f_j(i+\frac{\sigma_j}{V}), \quad \text{for } i = 1, ..., I-\frac{\sigma_j}{V}, \\ h_j(i) &= 0, \qquad\qquad\quad \text{for } i = I-\frac{\sigma_j}{V}+1, ..., I. \end{aligned}$$

Now $g_{j+1}(i)$ can be computed.

Per trip, the expected arrival delay $E\Delta_j$ is given by

$$E\Delta_j = \sum_{i=0}^{I} h_j(i) \cdot i \cdot V.$$

Finally the average expected delay over the different measuring points is calculated as

$$E\overline{\Delta} = \frac{1}{N}\sum_{j=1}^{N} E\Delta_j,$$

or a weighted average as

$$E\overline{\Delta_w} = \sum_{j=1}^{N} w_j \cdot E\Delta_j,$$

with w_j the weight for the arrival delay at the end of trip j and $\sum_{j=1}^{N} w_j = 1$.

Optimization To reach the optimal supplement allocation, a 2-OPT strategy is applied. The algorithm is started with some initial allocation of the supplements. The total amount of supplement should be the desired amount, but the initial allocation is not important, although the algorithm will find the optimal solution faster when the starting values are closer to the optimum.

For each pair of trips, it is evaluated whether an exchange of some supplement is beneficial. If the average delay decreases due to the supplement exchange, the supplements are adjusted, otherwise not. The comparison of supplement-exchanges is continued until no improvement can be found anymore.

To reach the optimal allocation faster, the comparisons are started with the exchange of a large amount of supplement. When no improvement can be found anymore, the exchange-size is decreased. This is repeated until the desired precision is reached. Note that an exchange-size smaller than the interval size V does not increase the precision of the solution. Furthermore, in the finally reached optimal supplement allocation, each trip can have one exchange-size supplement less or more than in the optimum. However, such small deviations are not of any practical importance. To reach the theoretical optimum, N-OPT exchanges for $N > 2$ can be used.

4.3.1 Numerical Results

The model described in Section 4.3 can be used for any number of trips and for any combination of disturbance distributions. This section shows the results for cases which are based on equally weighted and identically, exponentially, disturbed trips.

Two trips In the first case, a train run consisting of two trips is evaluated. Both trips are exponentially disturbed with an average of 1 minute. A total supplement of 2 minutes has to be divided over the 2 trips. This means that the total average incurred disturbances equal the total supplement. Due to the variability of the disturbances there will be a positive average delay. The optimum is equal to the optimum found by the analytical deduction in Section 4.2.1: $\sigma_1^* \approx 1.183$ and $\sigma_2^* \approx 0.817$, leading to $E\Delta_1 \approx 0.306$, $E\Delta_2 \approx 0.688$, and $E\overline{\Delta} \approx 0.497$. This solution was found by discretizing the disturbance distribution and delays into intervals of 0.001 minute. In each iteration 0.001 minute of supplement was transferred to another trip. A maximal delay of 20 minutes was taken into account.

Note that the problem is scalable: when increasing both the average disturbances and the total running time supplements by the same percentage, the optimal supplements for both trips are increased by that same percentage.

Looking at two trips and keeping the distributions exponential, we can compare the analytical results with the model computations for other total amounts of supplement. They appear to be the same. Here two additional results are given for the situation where both trips are exponentially disturbed with an average of 1. First for $\sigma_T = 4$ or twice the average disturbances, and then for $\sigma_T = 1$, or half the average disturbances. For $\sigma_T = 4$, $\sigma_1^* \approx 2.068$ and $\sigma_2^* \approx 1.932$, leading to $E\Delta_1 \approx 0.126$, $E\Delta_2 \approx 0.199$, and $E\overline{\Delta} \approx 0.162$. The increase of σ_T leads to a much more proportional allocation of supplements between the two trips. For $\sigma_T = 1$, $\sigma_1^* \approx 0.799$ and $\sigma_2^* \approx 0.201$, leading to $E\Delta_1 \approx 0.450$, $E\Delta_2 \approx 1.260$, and $E\overline{\Delta} \approx 0.855$. The supplements are now very much concentrated on the first trip.

When we compare these results with the case where $\sigma_1 = \sigma_2 = \frac{1}{2} \cdot \sigma_T$, the decrease of $E\overline{\Delta}$ is $3.1\%, 1.2\%$ and 0.2% for $\sigma_T = 1, 2$ and 4, respectively. So for two trips, the gain of redistributing the supplements does not seem to be substantial. However, in reliability every step forward is appreciated, and, as explained in the following intermezzo, a 3.1% average delay decrease leads to approximately a full percent-point increase of the 3-minute punctuality.

Punctuality gain It takes little mathematical effort to calculate the punctuality gain of a 3.1% decrease in average arrival delay. We make the assumption that the 3'-punctuality was 80% before the decrease and that arrival delays are exponentially distributed. This implies $1 - e^{-3\lambda} = 0.8$, with $\frac{1}{\lambda}$ the average delay size. Then $\ln(e^{-3\lambda}) = \ln(0.2)$ leads to $-3\lambda = -1.609$ or $\lambda = 0.536$ and an average delay of $1/0.536 = 1.864$ minutes.

A 3.1% decrease in average delay would then lead to an average delay of $1.864 \times 0.969 = 1.806$ minutes. This implies that the new $\lambda^* = 1/1.806 = 0.554$, and the 3'-punctuality equals $1 - e^{-3 \times 0.554} = 0.810$ or 81%. This is a gain of a full percent-point in the 3'-punctuality or a 5% decrease in the 3'-unpunctuality.

Total average disturbances equal to the total supplements In this section, the following problem has been solved: each trip of a train run is disturbed exponentially with average 1 and the total running time supplement equals the total number of trips, i.e. it is equal to the expected total average disturbances. Now allocate the total supplement to the trips, such that the average arrival delay of all intermediate points is minimal.

This question was answered by running the above model for a range of different numbers of trips. All these cases were discretized in intervals of 0.01 minute, and delays were cut off at 20 minutes. In each iteration, 0.01 minute of supplement was transferred to another trip. The results of the 10- and 25-trips cases are given in Figures 4.4 and 4.5. The vertical lines represent the *Weighted Average Distance*, or *WAD*. To calculate the relative distance of a supplement from the start of the line, the horizontal axis in Figure 4.6 (with the trips) is scaled such that the trips nicely fit into the 0-1 interval. The supplements are represented in the center of their respective trips. For example, for the case with 4 trips, the trips run from 0 to 0.25, from 0.25 to 0.5, from 0.5 to 0.75 and from 0.75 to 1. This leads to a representation of the accompanying optimal supplements at 0.125, 0.375, 0.625, and 0.875. In general terms, in the N-trip case, the j^{th} supplement is depicted at $\frac{2j-1}{2N}$ of the line.

Now we can define the *WAD* as the weighted average distance of the supplements from the starting point of the train line. It is calculated as

$$WAD = \sum_{j=1}^{N} \frac{2j-1}{2N} \cdot \sigma_j. \tag{4.9}$$

Not only for the 10-trip and 25-trip cases, but in all considered cases, with 2 to 25 trips, a similar shape of the supplement allocation was found. Collecting the results for 10 different cases, with 3, 4, 5, 6, 8, 10, 12, 15, 20, and 25 trips, results in Figure 4.6.

On the first trip, a relatively small supplement is added to the technically minimal running time. In Figure 4.6 this leads to the seemingly distinct observations on the left. After the first trip, the supplement increases to a level of around 1.35 minutes per trip. For the final trips the supplement rapidly decreases to 0.

Foremost, one has to realize that delay reductions do not only reduce the delay on the respective trip, but on all remaining trips. This means that a delay reduction is measured at all remaining measuring points. Consequently, a delay reduction early on weighs more heavily than a reduction later on, because it is measured more often. Only regarding this principle, one would expect to have most or all supplements very early on. But there is a counterweight: if there are no early delays, then early

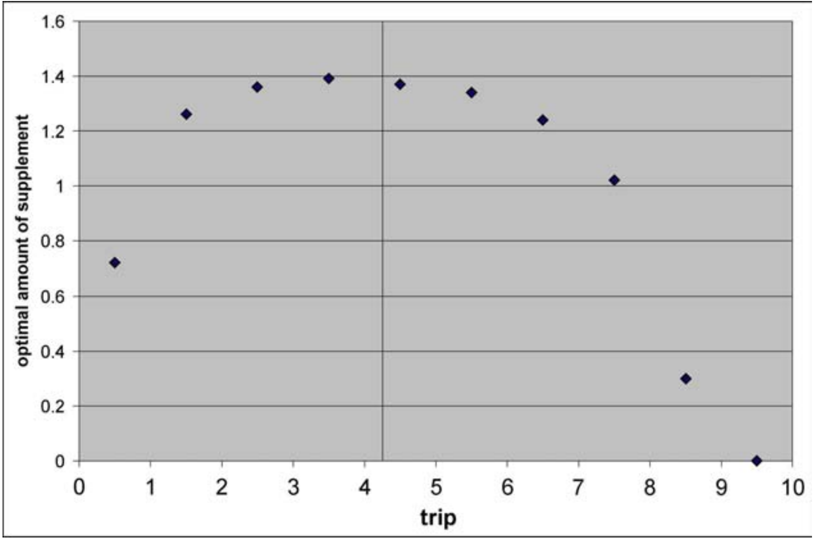


Figure 4.4: The optimal amount of running time supplement for 10 trips.

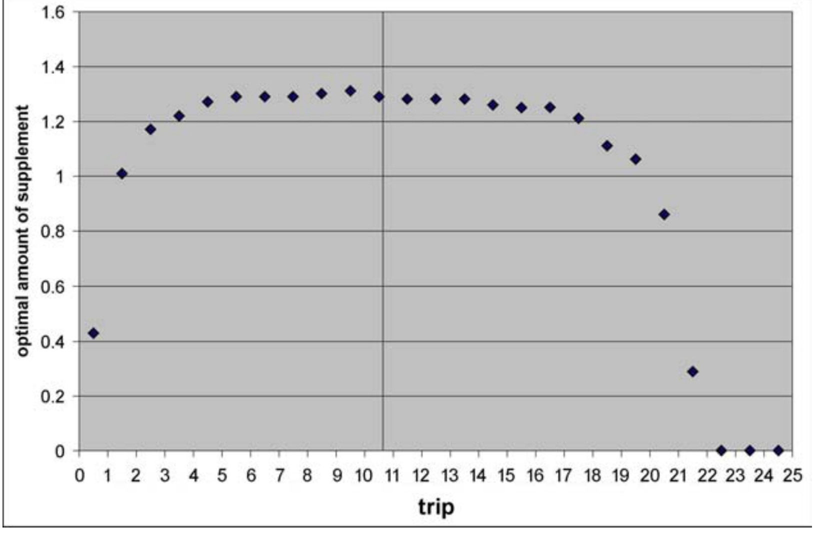


Figure 4.5: The optimal amount of running time supplement for 25 trips.

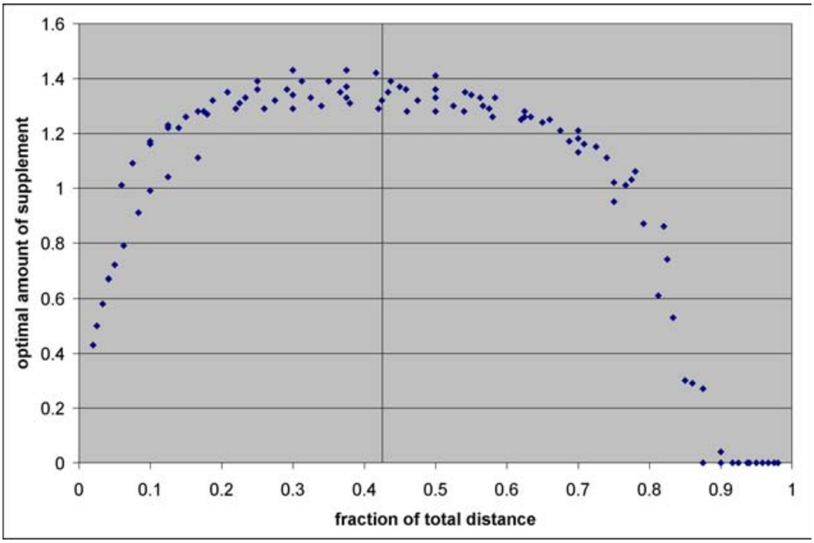


Figure 4.6: The optimal amount of running time supplement for all cases represented in one graph.

supplements are useless. These two reasonings together give a good explanation of the shape found in Figure 4.6.

As a consequence of the similar allocation of the supplements, the WAD is almost equal in all cases. To be more precise, it is close to 0.425.

Although the shape of the supplement allocation and the average location of the supplements is similar for any number of trips, the relative decrease in average delay increases with the number of trips. In Figure 4.7 the average delay of the optimal situation is compared to the situation with proportionally allocated supplements. The average delay decrease is only 1.2% for 2 trips, but the decrease is already 9.5% for 5 trips and 20.1% for 15 trips.

The optimal situations clearly show a large reduction in average delay compared to the situation with proportionally allocated supplements. However, during the optimization process it has also become clear, that small deviations in the supplement allocation hardly have any influence on the average arrival delay. The deviations in the objective value remain relatively small as long as there is little supplement on the first trip and no supplements on the last trips.

Used supplements The decrease in average delay is accomplished by a better allocation and a better utilization of the running time supplements. The total supplement usage in the optimal situation is less than in the case of a proportional

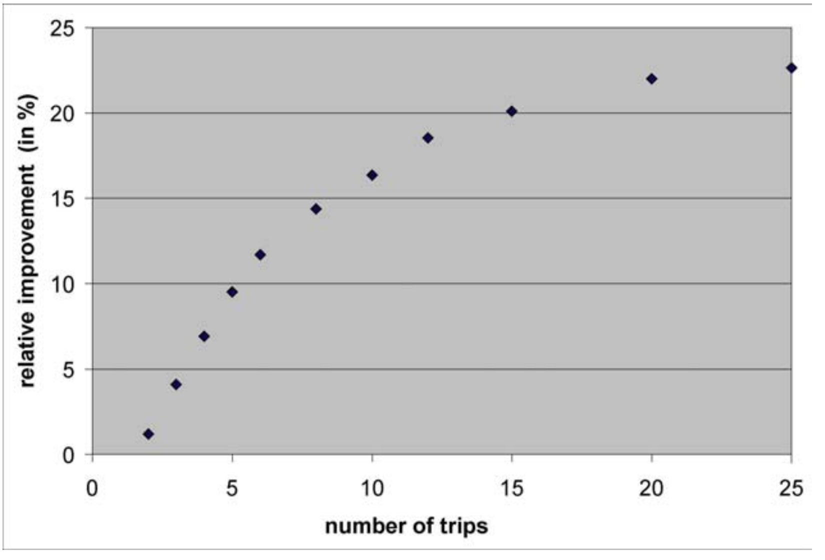


Figure 4.7: The decrease in average delay for the optimal supplement allocation as compared to a proportional allocation.

supplement allocation. To be precise, 72.0% of the supplements is used in case of the optimal supplement allocation, while 80.6% is used in case of a proportional allocation. Figure 4.8 shows the use of the supplements per trip for the cases with an optimal and proportional supplement allocation.

It may be counter-intuitive that the use of supplements in the optimal allocation is less than for the proportional allocation. The better performance in the optimal situation is accomplished by the earlier usage of these supplements. The early use of supplements implies an early delay decrease. That means that the smaller delay is measured more often than when the delay is decreased later on. This effect is made clear in Table 4.1 with the numerical outcomes of the case with ten trips. The average used supplements (in columns 5 and 9) decrease the average delay on all subsequent trips. In other words, the delay decrease on trip i accomplished by the supplements equals the cumulatively used supplements up to trip i (columns 6 and 10). The arrival delay at the end of trip i equals the cumulative disturbance (column 3) minus the cumulatively used supplements (columns 7 and 11). Thus, the optimal allocation of supplements is better than the proportional allocation, because the average cumulatively used supplements are larger.

Total average disturbances unequal to the total supplements In this section we look both at a case where the total average disturbances are double the

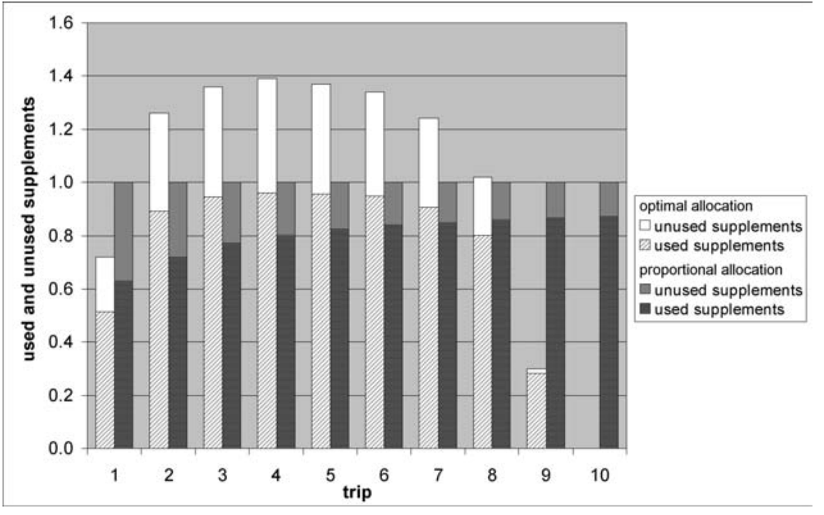


Figure 4.8: The used and unused supplements in the case with 10 trips.

Trip	Disturbance		Optimal Allocation				Proportional Allocation			
	average disturbance	cumulative disturbance	planned supplement	average used supplement	cum. used supplement	delay at end of the trip	planned supplement	average used supplement	cum. used supplement	delay at end of the trip
1	1.00	1.00	0.72	0.51	0.51	0.49	1.00	0.63	0.63	0.37
2	1.00	2.00	1.26	0.89	1.40	0.60	1.00	0.73	1.36	0.64
3	1.00	3.00	1.36	0.95	2.35	0.65	1.00	0.78	2.14	0.86
4	1.00	4.00	1.39	0.96	3.31	0.69	1.00	0.80	2.94	1.06
5	1.00	5.00	1.37	0.96	4.27	0.73	1.00	0.82	3.77	1.23
6	1.00	6.00	1.34	0.95	5.21	0.79	1.00	0.84	4.61	1.39
7	1.00	7.00	1.24	0.91	6.12	0.88	1.00	0.85	5.46	1.54
8	1.00	8.00	1.02	0.80	6.92	1.08	1.00	0.86	6.32	1.68
9	1.00	9.00	0.30	0.28	7.20	1.80	1.00	0.87	7.19	1.81
10	1.00	10.00	0.00	0.00	7.20	2.80	1.00	0.87	8.06	1.94
Avg.	1.00	5.50	1.00	0.72	4.45	1.05	1.00	0.81	4.25	1.25

Table 4.1: The used supplements and their effectiveness.

size of the total supplements, and at a case where the disturbances are half this size.

The first case that is presented here is a case with 10 trips, where each trip is still exponentially disturbed, but now with an average of 2 minutes. There is, on average, still 1 minute of supplement available per trip.

In these new cases, a discretization interval of 0.01 minute was used again, and delays were cut off at 20 minutes. In each iteration 0.01 minute of supplement was transferred to another trip.

The optimal supplement allocation is now more concentrated on the earlier trips. This is understandable, since early supplements are still weighted more often, but the probability for excessive supplements decreases when the disturbances increase. The shift of the supplements to the left is quantified by the WAD which decreases from 0.42 to 0.32. The delay reduction is slightly larger than in the case where the supplements equal the average disturbances.

When considering the opposite situation, where the total average disturbances are only half as large as the total supplements, the results also shift into the opposite direction. The weighted average distance of the supplements approaches 0.5 (0.492). The delay decrease with respect to the case with proportionally allocated supplements is about 3%.

The results for the case with 10 trips and 5, 10 and 20 minutes of total disturbances are shown in Figure 4.9. The optimal supplement allocation obviously depends on the disturbance distribution.

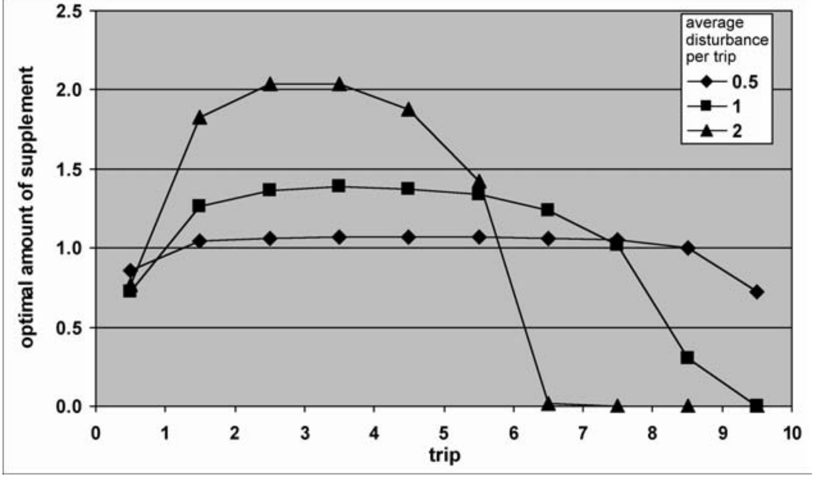


Figure 4.9: The optimal supplement allocation for different disturbances.

4.3.2 Alternative Objectives

The optimizations sofar were all based on the same objective: a minimal average arrival delay. Harrod (2003) introduces a threshold, below which delays are considered 'non-objectionable'. In line with this threshold, he uses two other objectives: minimize the number of passenger arrivals with a delay larger than the threshold, and minimize the passenger arrival delays in excess of the threshold. In fact, the first objective mentioned above is the same as maximizing the (weighted) punctuality.

The model described in Section 4.3 is also able to optimize with respect to the punctuality, and the average excess delay above the threshold can also be handled.

To investigate the influence of the threshold on the optimal supplement allocation, ten trips are regarded. Each of the trips incurs an exponentially distributed disturbance with an average of 1 minute. The total supplement for the ten trips is 10 minutes.

First, unpunctuality was minimized with thresholds of 1, 2, 3, 4 and 5 minutes. Again, a discretization interval of 0.01 minute was used, and delays were cut off at 20 minutes. In each iteration 0.01 minute of supplement was transferred from one trip to another. The optimal supplement allocation for these cases is depicted in Figure 4.10. The shape is comparable with the shape in Figures 4.4 and 4.6. However, some features of the optimal supplement allocation for punctuality maximization with different thresholds have to be mentioned.

First of all, the amount of supplement on the first trip decreases when the threshold increases. This can be explained by the fact that small delays are not taken into account in large-threshold measurements. Secondly, the differences in supplement on the first trip are compensated almost only by trips 6 and 7: here the large-threshold cases have the largest supplement. Finally, note that in the 5-minute threshold situation, the optimal amount of supplement is larger on the seventh than on the sixth trip. We are unable to explain this phenomenon.

The objective values for the supplement allocations in Figure 4.10 are shown in Table 4.2. The punctuality increase which can be reached by optimizing the supplement allocation (compared to a proportional allocation) increases with the size of the threshold. Still, the weighted average distance of the supplements goes towards 0.5 when the threshold goes up.

The average excess delay was minimized for the same thresholds. The discretization intervals were 0.01 minute in these cases also. Delays were cut off at 20 minutes again, and 0.01 minute of supplement was transferred from one trip to another in each iteration. In addition, the case with no threshold is included in Figure 4.11. This corresponds to the 'regular' situation where average delay is minimized.

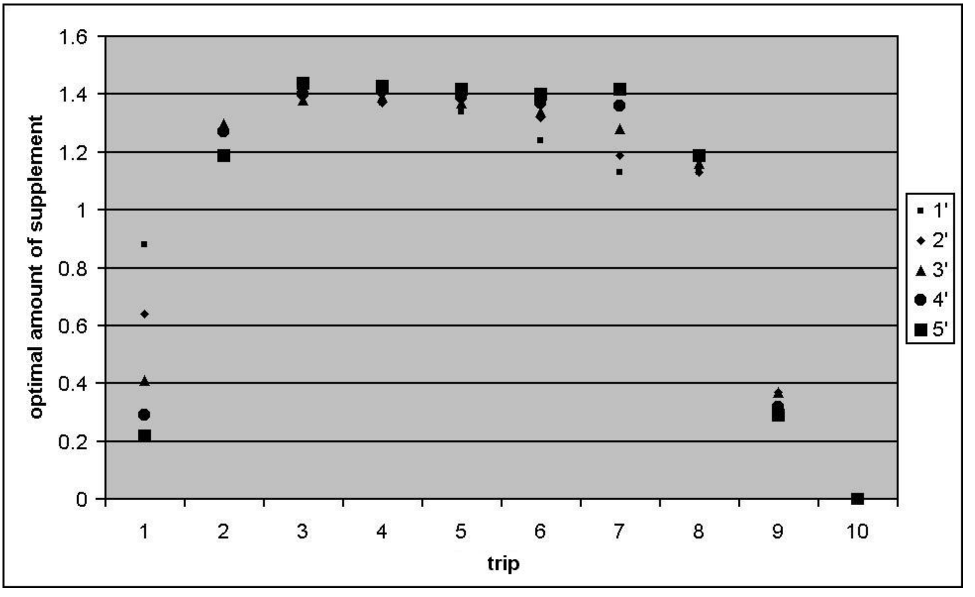


Figure 4.10: Supplement allocation in case of punctuality maximization for different thresholds.

threshold in minutes	unpunctuality in %		unpunctuality decrease (%)	*WAD* for the optimum
	for a proportional allocation	for an optimal allocation		
1	39.54	33.93	14.2	0.417
2	23.94	19.61	18.1	0.432
3	14.18	11.06	22.0	0.442
4	8.23	6.10	25.8	0.448
5	4.68	3.31	29.4	0.452

Table 4.2: Unpunctuality decrease compared with a proportional allocation for different thresholds.

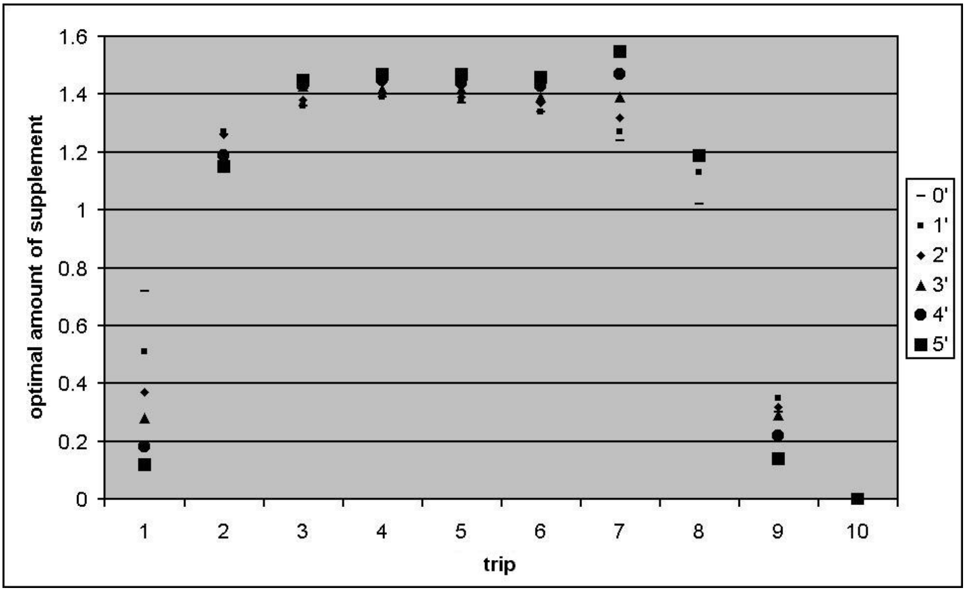


Figure 4.11: Supplement allocation in case of excess delay minimization for different thresholds.

The remarks made about Figure 4.10 can also be made here. The increasing amount of supplement for the seventh trip is now more dominant and also true for the 4-minute case. Furthermore, note that on the ninth trip the amount of supplement is reversed again, and the supplements on the ninth trip are largest for the cases with the smallest thresholds.

The average excess delays for a proportional and optimal supplement allocation are presented in Table 4.3.

4.3.3 The Numerical Model Evaluated

The model described in this section provides a good insight into the optimal allocation of supplements. Some important properties of the model are summarized below.

- **Discretization** The model uses small intervals to evaluate the delay propagation over a train line. This implies that the disturbance distribution has to be discrete (or discretized) as well. However, any kind of discrete disturbance distribution can be used in the model. Moreover, different distributions can be used for different trips.

threshold in minutes	excess delay in minutes: for a proportional allocation	excess delay in minutes: for the optimal allocation	delay decrease (%)	WAD for the optimum
0	1.253	1.049	16.3	0.423
1	0.747	0.596	20.2	0.437
2	0.436	0.332	23.9	0.444
3	0.250	0.181	27.5	0.449
4	0.141	0.098	30.6	0.452
5	0.078	0.052	33.3	0.454

Table 4.3: Excess delay decrease for different thresholds.

- **Model running time** The running time of the model presented in this section mainly depends on three factors. First of all the number of trips is important, since the possible pairwise exchange of supplements is quadratic in the number of trips. Secondly, the number of discretization intervals I has a large influence, since the model multiplies $I \times I$-matrices to determine the delays. Furthermore, starting values of the supplements close to the optimal allocation decrease the running time considerably. For two trips, 2000 intervals (i.e. discretization in 0.01 minute intervals and a maximum delay of 20 minutes), and a proportional starting allocation of the supplements, the running time is a few minutes. For 25 trips, ceteris paribus, the model has to run for a day.

- **Flat objective function** The average arrival delay over all trips is relatively insensitive to small deviations in the supplements. The transition of a small amount of supplement to another trip often changes the objective with a much smaller fraction. Because of the discretization of the disturbance distributions, this can lead to ‘optimal’ solutions which deviate from the ‘real’ optimum for the continuous disturbance distribution.

- **No secondary delays** The model only regards primary delays and their propagation on the same train line. Secondary delays from and to other lines are left out.

- **Only one line** Because only one line is considered by the model, it is not able to observe any influences for an entire network timetable.

- **Departure delay** Despite the use of the assumption, that the departures from station 0 are not delayed, the model can handle any finite discretized distribution for this first departure.

- **Objective functions** Due to the fact that the objective function is only used to compare possible outcomes of the model, it can be of any mathematical form and is not restricted to linearity.

In the next section, the model is applied to determine the optimal supplement allocation for the real life corridor Haarlem–Maastricht/Heerlen.

4.4 A Practical Case

To support the theoretical results from Sections 4.2 and 4.3, a practical simulation case has been worked out. Two long lines (the 800 from Haarlem to Maastricht and the 900 from Haarlem to Heerlen) are the backbone of the case. In practice, the arrival times of each train in the 800- or 900-line are measured eight times per direction. The distance from Haarlem to Maastricht or Heerlen is close to 250 km, which is covered in just under three hours.

4.4.1 Case Description

The case includes all passenger trains on the tracks from Haarlem (Hlm) via Amsterdam Central (Asd), Utrecht Central (Ut) and Eindhoven (Ehv) to Maastricht (Mt) and Heerlen (Hrl). All these lines are depicted in Figure 4.12. The international train, the long distance trains, the interregional trains and short distance trains are represented by double (═══), solid (───), broken (— — —), and dotted (·········) lines, respectively. Each line represents one train per hour.

The timetable is executed on a double track, one for each direction. In Haarlem, Amsterdam Central, Abcoude (Ac), Utrecht Central, Geldermalsen (Gdm), 's-Hertogenbosch (Ht), Eindhoven, Roermond (Rm), Sittard (Std), Maastricht and Heerlen, more tracks are available for starting and ending lines, and for overtaking. The train lines and train order are identical to those of the 2004 NS rush hour timetable. This includes overtaking in Abcoude, Geldermalsen and 's-Hertogenbosch twice per hour per direction. The resulting timetable, given in Appendix B is not exactly the same as the NS-timetable, but the structure is comparable.

The simulated timetable was constructed with the timetabling tool DONS. The running time supplements were adjusted for both situations. Note that for the different situations, the planned running times were kept the same. Varying the running

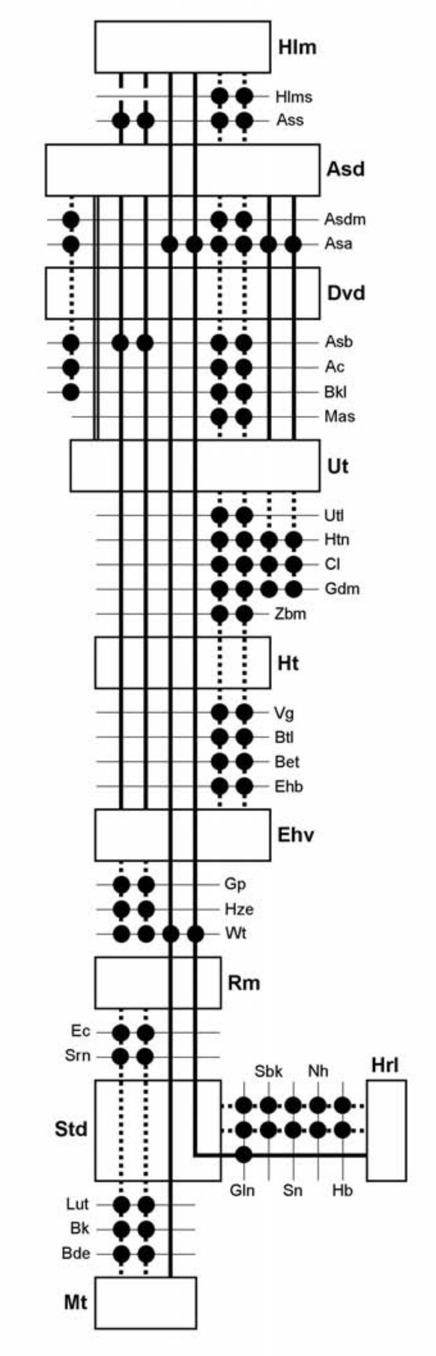


Figure 4.12: Train lines and dwelling patterns of the practical case. All abbreviations indicate a station; delays are measured at the rectangular stations.

time supplements implies that the assumed technically minimal running times differ between the situations. This would be exactly the other way around in reality, where one has to deal with a fixed technically minimal running time and varying timetabled running times.

The average running time supplements in the original DONS timetable are 7.92% on top of the technically minimal running times. No dwell time supplements were timetabled, except for the trains which are overtaken. Exceptional running time supplements of five minutes are included in the running time of the long distance train from Amsterdam Central to Eindhoven between Duivendrecht and Utrecht Central, and also in the opposite direction. This line is slowed down in the timetable because of a local train on the same track.

The large stations, where the delays are measured, such as depicted in Figure 4.12, are now stations 0 up to 8 as in Figure 4.1: train lines 800 and 900 consist of 8 trips each. For the simulations, all running time supplements are situated just before these large stations.

The optimization model in Section 4.3 was used to determine the optimized situation. Only the supplements for the 800- and 900-lines were reallocated; the other lines keep the same supplements as in the reference situation. This is mainly based on the results from the theoretical cases which showed that punctuality gains are small for lines with only a few trips.

4.4.2 Experimental Design

The optimization of the running time supplement allocations is based on exponential disturbances with an average equal to the running time supplements from the reference case. This is comparable to the theoretical cases in Section 4.3, with different average delays for different trips. These disturbances are also applied in the simulations.

The optimization has been performed four times: both the 800- and 900-lines northbound and southbound. This is because the disturbances for these lines, and therefore also the optimal supplement allocations, are different. As in the theoretical cases, a discretization interval of 0.01 minute was used, and the maximal delay was 20 minutes. In each iteration 0.01 minute of supplement was transferred to another trip.

4.4.3 Results

The results of the optimizations are presented in Table 4.4.

trip	planned running time (min)	average disturbance (min)	running time supplements (min) proportional 800 & 900	optimal 800	optimal 900
Hlm-Asd	14	1.03	1.03	0.85	0.87
Asd-Dvd	11	0.81	0.81	1.02	1.03
Dvd-Ut	17	1.25	1.25	1.46	1.48
Ut-Ht	28	2.05	2.05	2.60	2.65
Ht-Ehv	18	1.32	1.32	1.70	1.74
Ehv-Rm	31	2.27	2.27	2.55	2.55
Rm-Std	15	1.10	1.10	0.75	0.83
Std-Mt	15	1.10	1.10	0.00	-
Std-Hrl	18	1.32	1.32	-	0.00
Hrl-Std	18	1.32	1.32	-	1.00
Mt-Std	16	1.17	1.17	0.86	-
Std-Rm	15	1.10	1.10	1.19	1.19
Rm-Ehv	30	2.20	2.20	2.69	2.73
Ehv-Ht	19	1.39	1.39	1.88	1.88
Ht-Ut	28	2.05	2.05	2.55	2.55
Ut-Dvd	18	1.32	1.32	1.50	1.47
Dvd-Asd	12	0.88	0.88	0.47	0.47
Asd-Hlm	14	1.03	1.03	0.00	0.00

Table 4.4: The running time supplements for the proportional allocation (7.92% of the minimal running time) and the optimized situation. The top half represents the southbound trains, the bottom half the northbound trains.

The table shows that, in line with the results for trips with equal disturbances, the supplements are moved towards the start of the line in comparison with a proportional allocation. To calculate the *Weighted Average Distance* or *WAD*, the distance is now expressed relatively with respect to the technically minimal running time. This is in line with the disturbances, because these are also a fraction (7.92%) of the technically minimal running time. This leads to the following equation:

$$WAD = \sum_{j=1}^{8} \frac{(\sum_{k=1}^{j-1} r_k) + \frac{1}{2} r_j}{\sum_{k=1}^{8} r_k} \cdot \sigma_j, \tag{4.10}$$

where r_j and σ_j are the technically minimal running time and the running time supplement, respectively, for trip j. Note that (4.10) is equivalent to (4.9) when

$r_j = r_{j'}$ for all $j \neq j'$ and $j, j' \in 1, ..., 8$. Now the WAD for the reference cases, where the average disturbance is equal to the running time supplement on the respective trip, is 0.5. For the optimal supplement allocations, the WAD is between 0.431 and 0.454.

The optimal supplement allocation for the 800- and 900-lines decreases the arrival delays at the measuring points with 9.8% to 10.8% with respect to the reference case. However, this result includes only primary delays. The question is which influence the reallocation of supplements has on the delay propagation. To investigate this question a SIMONE simulation has been carried out.

In the simulations, every train incurs a 'relative' disturbance between each pair of large stations. These disturbances are randomly picked from an exponential distribution with an average equal to 7.34% of the planned running time. When the running time supplements are distributed relative to the running times, this is 7.92% of the technically minimal running time.

In line with the optimization model, the simulations showed a sizable decrease in average delay. The average arrival delay of the 800- and 900 lines decreased from 3.15 minutes to 2.69 minutes: an improvement of 14.70%. This is more than the model results without secondary delays. Apparently, the 800- and 900-lines incur less secondary delays when they are more punctual themselves.

Figure 4.13 shows the delays of the 800- and 900-lines in southern direction. At the bottom one can see the model results: the small squares represent the average arrival delay in the numerical model for the reference case; the small diamonds show the delays for the optimal supplement allocation. The larger squares and diamonds represent the simulation results for the two situations. These simulation results include secondary delays, which are not present in the model of Section 4.3. The same representation for the northern direction is given in Figure 4.14.

Most secondary delays are incurred if a fast intercity train leaves a large station just behind a local train, and is not able to overtake it. In that case, the fast train incurs several minutes of secondary delay for each small station that is passed. This happens most frequently when an intercity train has a sizable delay before leaving a large station, where it is timetabled just before a local train. The highest density of local trains can be found between Geldermalsen and Amsterdam Central. For the 800- and 900-lines, the northbound passing of Geldermalsen and the departure from Utrecht Central are very critical, being planned just before a local train. Note that the 800- and 900-trains have already traveled for almost two hours before passing Geldermalsen and over two hours before reaching Utrecht Central. Therefore, they may be heavily disturbed already. The same holds for Roermond in southern direction, but the consequences are smaller, because the (planned) running time differ-

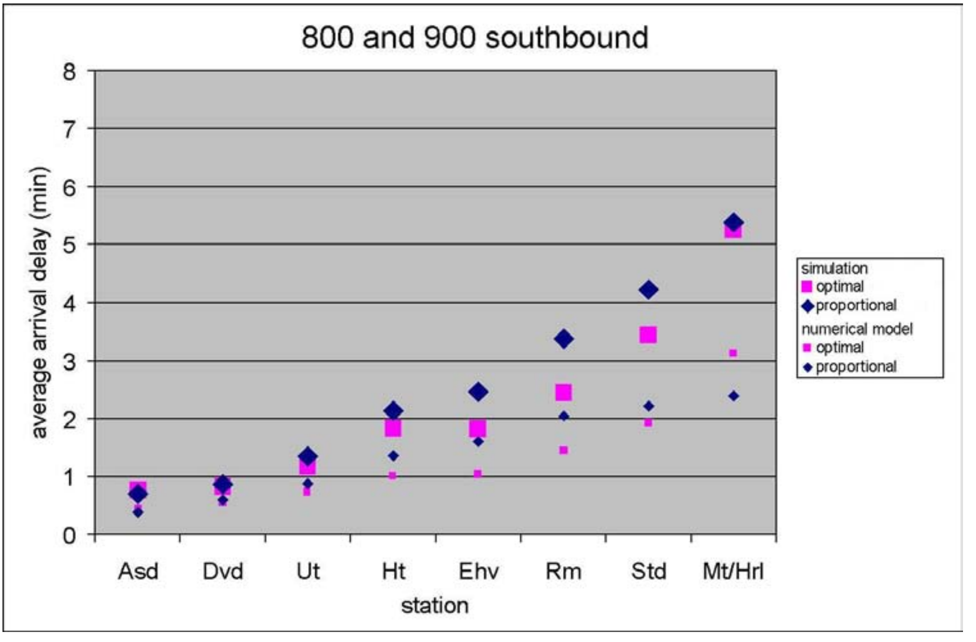


Figure 4.13: Results for the 800- and 900-lines Southbound. The simulation results for the optimal (large diamonds) and reference (large squares) situations. The model results, not including secondary delays, are also given for both situations.

ences between the intercity and local train are smaller. This explains the large delay increase, for both situations, between Geldermalsen, Utrecht Central, and Duivendrecht in northern direction. However, the increase for the optimized situation is more moderate, since the expected delays before reaching Geldermalsen and Utrecht Central are smaller.

The large increases in delay at the last trips (Sittard-Heerlen/Maastricht southbound, and Amsterdam Central-Haarlem northbound) in the optimized situations are due to the absence of any running time supplements on these trips.

Besides the better performance of the 800- and 900-lines, the other lines are also doing better: the average arrival delay of all other lines decreased by 3.25%. This implies that the optimized and more punctual 800- and 900-lines also cause less secondary delays to the other trains than in the reference situation. A small summary of the decrease in delays is given in Table 4.5.

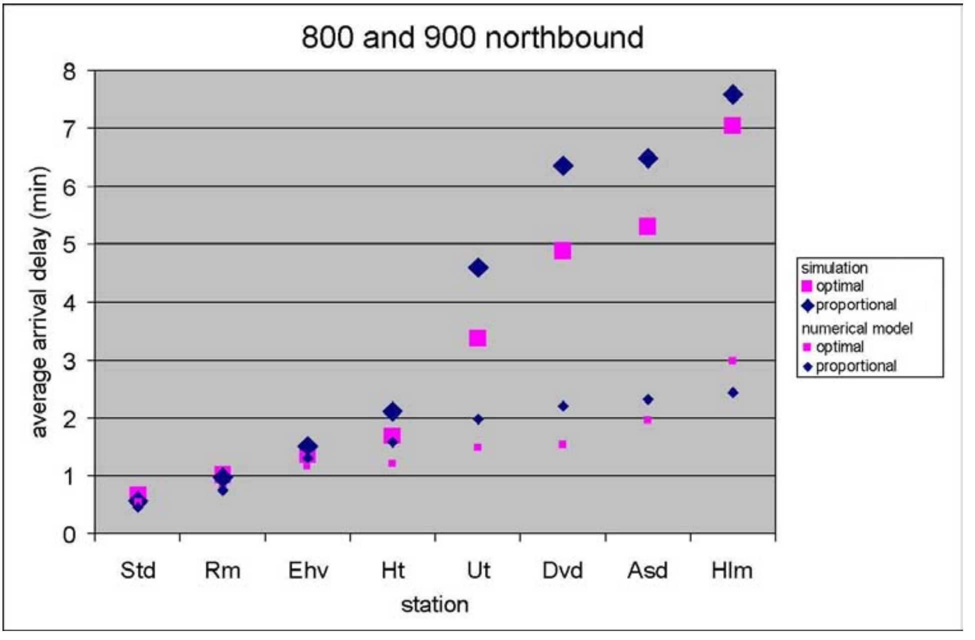


Figure 4.14: Results for the 800- and 900-lines Northbound. The simulation results for the optimal (large diamonds) and reference (large squares) situations. The model results (not including secondary delays) are also given for both situations.

4.5 Conclusions and Discussion

4.5.1 Results

The results in this chapter indicate that a proportional allocation of the running time supplements does not lead to a minimal average delay. The supplements on the earliest and especially the last trips have to be below the average supplement as is shown in Figure 4.6.

The supplements on the earlier trips have to be relatively small, because the incurred delay of the train at the start of its journey is still small as well. The

	southbound	northbound	south & north
800 and 900	12.46%	16.20%	14.70%
other lines	2.56%	3.66%	3.25%
all lines	8.44%	10.79%	9.87%

Table 4.5: Decrease in average arrival delay for the different directions. The table indicates the relative gain of the optimized situation compared to the reference case.

supplements on the earlier trips which are not used to decrease the delay are lost.

The supplements on the last trips decrease the delay on the last trips only. This decrease is now only measured at the last arrival station. Delay decreases, which are realized by supplements earlier on, will be measured several times. Early supplements, which are smaller than the delays, have therefore a larger influence on the average delay than later supplements. Indeed, if it would be certain that the total supplement can be used to decrease the delay on the first trip, then all supplements should be concentrated on the first trip.

Figure 4.6 illustrates that the allocation of the supplements is almost insensitive to the number of trips. When the total expected disturbances are equal to the total amount of supplement, there is no supplement on the last trips, and in the middle part supplements are about 30% above the average. The delay decrease is 1.2% for two trips, but already 10% for five trips and 20% for fifteen trips.

When the disturbances are relatively large compared to the total supplements, a relatively larger portion of the supplements is moved towards the earlier trips. This is because the probability that supplements remain unused is small. The difference in average delay between the optimal supplement allocation and the proportional supplement allocation increases. The opposite is true for relatively small disturbances.

4.5.2 Practical Considerations

Reallocation of supplements does not only influence the average delay. A few other factors are discussed here.

- **Rounding** In practice, departure and arrival times are rounded to minutes. However, the supplement shifts indicated by the optimization in this section concern at most a few tenths of a minute. So exactly realizing the optimal supplement allocation in the timetable may be impossible. Still, the reallocation can imply that roundings will be downwards instead of upwards, or the other way around.

- **Running time differences** A first trip for one train may be on the same open track as the last trip for another train. This implies that some trains receive less than average supplements, while other trains receive above average supplements on the same open track. This leads to other running time differences between trains. Especially when running time differences increase, timetabling possibilities decrease, or it is even impossible to find a timetable. As will be explained in Chapter 5, large running time differences also lead to a punctuality decrease. An integrated approach is to be preferred, see Chapter 6.

- **Symmetry** The symmetric nature of a timetable depends on 'symmetric' running times: for each trip the running time must be equal for both directions of a train line. When supplements are reallocated, this may break the symmetry of the timetable. However, the size of the shifts in supplements are so small, that the practical impact on the symmetry is negligible.

Chapter 5

Heterogeneity of Railway Traffic

Railway traffic is considered homogeneous if all trains have similar characteristics; especially the same speed, resulting from running times and stopping times. Good examples of homogeneous railway traffic are metro systems where all trains have the same running times and stop at all stations. However, for national railway networks, railway traffic cannot be fully homogeneous. Usually cargo trains and passenger trains share the same infrastructure. But probably more important, there is a large differentiation in passenger services, ranging from short distance trains (which dwell at all stations underway) to international high speed connections (with high speeds, only stopping at a few large stations), partly sharing the same infrastructure. When there are large differences in timetable characteristics for trains on the same track, the railway traffic is called heterogeneous.

Heterogeneous railway traffic leads to small headways in the timetable. These small headways tend to increase the delay propagation. Therefore we analyze the relation between the heterogeneity and the reliability of timetables in this chapter.

After an introduction on homogenization in Section 5.1, two new heterogeneity measures, SSHR and SAHR, are introduced in Section 5.2. A theoretical case in Section 5.3 and a practical case in Section 5.4 do not only show the impact of speed differences on delay propagation, but they also point out the usefulness of the SSHR and SAHR for quantification of the heterogeneity problem. Some additional consequences of homogenization are discussed in Section 5.5. Section 5.6 gives some concluding remarks.

5.1 Homogenization

Homogenization of a railway system means that differences in running times of different trains along a railway line are decreased. There are several alternative options for homogenization:

- **Slowing down intercity trains** Decreasing the speed of an intercity train means longer running times for these, usually considered more prestigious, services. On the other hand, besides the homogenization effect, the extra running time supplement created in this way will increase reliability.
- **Speeding up local or stopping services** Decreasing running times can only be achieved by decreasing running time supplements or by using faster rolling stock. The first option can be very hurtful for reliability and is very restricted in size, the second option is probably very costly.
- **Overtaking** When slower services are overtaken by faster services, running time differences should only be regarded from or up to this overtaking station. An important prerequisite is the presence of an overtaking track: a second track for the same direction is needed. One of the disadvantages of overtaking is the interdependency between both trains at the overtaking station. It also leads to a time loss for the stopping service.
- **Shorter lines for the stopping services** By decreasing the length of stopping services, stopping and intercity services share the same infrastructure for shorter distances and the difference in number of stops decreases. Unfortunately, this leads to more passenger transfers. In theory, shorter stopping services have almost the same effect as overtaking, if the shorter services have the overtaking stations as start and ending points. The difference is found in the rolling stock circulations. At first glance, shorter services lead to less dependencies in the network, but turning around at line-endpoints can also lead to additional delays and additional conflicting routes.
- **Equalizing the numbers of stops** Adding some stops to the intercity services leads to smaller differences between services. The small stations which are now serviced by the intercity trains, can be skipped by the stopping services. When repeating this until the number of stops of the stopping trains is equal to that of the intercity services, a maximal homogenization can be reached. One cannot speak about stopping trains and intercity trains anymore! A very harsh way to equalize the number of stops is to close down some minor stations. These could be serviced by busses.

In this chapter, the last option is chosen: we equalize the number of stops to create more homogeneous timetables. Using different amounts of running time supplements or different kinds of rolling stock makes comparisons between heterogeneous and homogeneous timetables unfair. This discards the options of slowing down intercity trains or speeding up local trains. Additional overtaking or shortening local services is often impossible without large investments for new sidings, where trains can be overtaken or turn around. The final option, equalizing the number of stops, seems the easiest to implement on the short term and is, maybe together with additional overtakings, the most promising on the long term.

When heterogeneous services share the same infrastructure over large distances, timetabling becomes very complicated. Heterogeneity usually leads to many small headway times, which increases delay propagation in the operations. This chapter provides a scientific analysis of this problem area. To that end, we first develop two heterogeneity measures. Then simulation of both theoretical and practical cases is used to show the importance of homogeneity of a timetable. Besides the fact that we use the heterogeneity measures to compare different timetables, it is also intended to be useful for the development of timetables for real world operations. In this chapter we assume cyclic timetables again.

5.2 Heterogeneity Measures

The most commonly used measure for line capacity in railways is capacity consumption. This measure is described in UIC Leaflet 406 (UIC, 2004). For any chosen time interval it measures the percentage of time necessary to operate the planned trains at technically minimal headway and at technically minimal running times. The disadvantage is that this measure does not make a difference between some very distinctive situations. For example, the individual headways do not play a role in the capacity consumption, only the total of the headways. This means that the capacity consumptions of the two timetables in Figure 5.1 are the same. Furthermore, the capacity consumption is also indistinctive between a situation with thirty trains per hour, and the situation in Figure 5.1(b) (given that the minimal headway is two minutes). Therefore we attempt to develop new measures based on headways.

Given the frequency of a line, the average headway at a location along that line is simply equal to the cycle time divided by the frequency. More useful headway measures are described by Carey (1999). He shows that equalizing scheduled headways for one station has a positive influence on punctuality, when disturbance distributions are sloping downward and are equal for all trains. The measures he describes are based on this principle. These include

- the percentage of headways smaller than a certain size;
- the percentiles of the headway distribution;
- the range, standard deviation, variance, or mean absolute deviation of the headways.

The further description of these measures implies that the headways are measured at one single location.

Above, several possible measures with respect to the spread of headways are given. However, the measures with percentages or percentiles only regard part of the headways, in particular the smallest. But in the Netherlands even the largest headways are small enough to be considered. The other measures, knowingly range, standard deviation, variation and mean absolute deviation of headways, can only be used for a given number of trains per time interval. Furthermore, all these measures focus on one specific location only, and not on a track.

An important disadvantage of measuring headways at only one location is that it does not tell anything about the train behavior on the surrounding tracks. Therefore, we take the smallest headways between two consecutive trains on a certain track section instead of at one certain location. When all trains on a certain track section are one hundred percent homogeneous, the sum of the smallest headways on this track section is equal to the cycle time. When traffic on a certain track is very heterogeneous, the short distance trains depart just after the long distance trains at the start of the track section, and the long distance trains arrive just after the short distance trains at the end of the track section, leading to a small total sum of smallest headways.

The disadvantage of just taking the sum of the smallest headways in a linear way is that it does not take into account how the trains are spread over the cycle time. With a cycle time of sixty minutes with four homogeneous trains, one will always have a total sum of (smallest) headways of 60', whether these trains are nicely spread, such as in Figure 5.1(a) with four 15-minute intervals, or not, as in Figure 5.1(b) with headways of 5, 25, 5 and 25 minutes, respectively. However, taking the sum of reciprocals gives a clear distinction between these situations. In particular, the examples in Figure 5.1 lead to $\frac{1}{15}+\frac{1}{15}+\frac{1}{15}+\frac{1}{15} \approx 0.27$, and $\frac{1}{5}+\frac{1}{25}+\frac{1}{5}+\frac{1}{25} \approx 0.48$, respectively.

This leads to our first heterogeneity measure, based on both the heterogeneity and the spread of trains over the hour. This measure is applicable to railway tracks

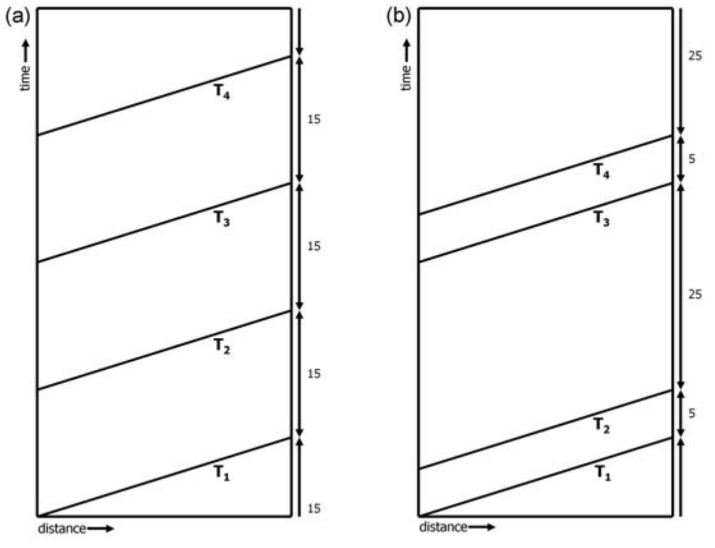


Figure 5.1: Time-distance diagram for two homogeneous situations with a different headway distribution.

between two neighboring railway nodes. For N trains per cycle, the Sum of Shortest Headway Reciprocals (SSHR) is defined as:

$$\text{SSHR} = \sum_{i=1}^{N} \frac{1}{h_i^-} \tag{5.1}$$

with h_i^- the smallest scheduled headway between trains i and $i+1$ on the track section, and train N is followed by train 1, due to the cyclicity of the timetable.

As stated earlier, the SSHR is not only capable of representing the distribution of trains over the hour on a track, but also of including the heterogeneity of these trains on this track. The homogeneous situation in Figure 5.1 gives an SSHR of 0.27. The slightly heterogeneous situation in Figure 5.2(a) leads to an SSHR of $\frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} \approx 0.44$. Figure 5.2(b) represents a very heterogeneous situation with an SSHR of $\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 2$.

A disadvantage of the SSHR is that headways at departure are penalized as heavily as headways at arrival. However, in practice headways at arrival seem to be more important than headways at departure. The first reason is that delays at arrival are, on average, larger than at departure. Secondly, faster long distance trains can be caught behind short distance trains towards the end of a railway section, which implies a sizable secondary delay. Therefore we developed a second measure, which

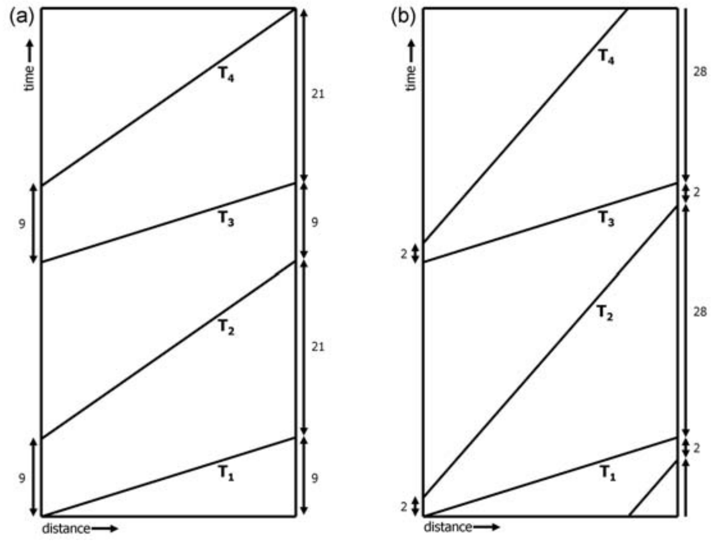


Figure 5.2: Two time-distance diagrams with slightly heterogeneous (a) and very heterogeneous (b) railway traffic.

only depends on the arrival headways between every pair of subsequent trains, the Sum of Arrival Headway Reciprocals (SAHR):

$$\text{SAHR} = \sum_{i=1}^{N} \frac{1}{h_i^A} \tag{5.2}$$

with h_i^A the scheduled headway at arrival between trains i and $i+1$.

In homogeneous cases, the SAHR is equal to the SSHR, so the SAHR is 0.27 in Figure 5.1(a) and 0.48 in Figure 5.1(b). In a specific heterogeneous case, the SAHR is always smaller than the SSHR. The timetables represented in Figure 5.2 have an SAHR of $\frac{1}{9} + \frac{1}{21} + \frac{1}{9} + \frac{1}{21} \approx 0.32$ and $\frac{1}{2} + \frac{1}{28} + \frac{1}{2} + \frac{1}{28} \approx 1.07$, respectively.

Unfortunately, the SAHR does not take the track into account anymore and is in fact a single location measure. Still, the arrivals can only be evenly spread over the hour if the timetable is not too heterogeneous. This means that heterogeneity is implicitly taken into account. However, an improved measure may be attained by taking the weighted average of the two measures above.

The two measures developed above are not absolute measures, but are mainly meant to be able to compare different timetables for the same track or as an indication of how to produce a reliable timetable for a certain track.

Minimal headways The new heterogeneity measures SSHR and SAHR were discussed without referring to the technically minimal headways. However, during the operations, it does not matter what the absolute size of a headway is, but what the buffer in this headway is: the difference between the planned headway and the minimal headway. Using headway buffers instead of planned headways for the SSHR would lead to the Sum of Shortest Buffer Reciprocals (SSBR) with the following formula:

$$\mathrm{SSBR} = \sum_{i=1}^{N} \frac{1}{(h_i - h_i^{min})^-} = \sum_{i=1}^{N} \frac{1}{b_i^-} \tag{5.3}$$

where $(h_i - h_i^{min})^-$ is the smallest difference between the planned headway and the minimal headway between trains i and $i+1$. This equals the minimal buffer time b_i^- between these two trains.

Although this measure seems more fair, one or two small buffers will increase the SSBR tremendously, because the reciprocal of the buffer tends to infinity when the buffer goes to zero. The difference between a *small* and a *very small* headway would be too large. The SSBR even becomes useless when the headway buffer is zero or even slightly negative.

A way of increasing the applicability of the SSBR is increasing the denominator of all i terms with a fixed number q. Let us call the resulting measure the Adjusted SSBR (ASSBR):

$$\mathrm{ASSBR(q)} = \sum_{i=1}^{N} \frac{1}{b_i^- + q} \tag{5.4}$$

A natural choice for q would be the average minimal headway $\overline{h^{min}}$, leading to the following:

$$\mathrm{ASSBR}(\overline{\mathrm{h^{min}}}) = \sum_{i=1}^{N} \frac{1}{b_i^- + \overline{h^{min}}} = \sum_{i=1}^{N} \frac{1}{(h_i - h_i^{min})^- + \overline{h^{min}}} \tag{5.5}$$

Indeed, under the assumption that the minimal headway is the same everywhere, we find that $\mathrm{ASSBR}(\overline{\mathrm{h^{min}}}) = \mathrm{SSHR}$. In fact, in the remainder of this chapter we assume a minimal headway of two minutes everywhere.

A similar reasoning holds for the SAHR and the related SABR and ASABR(q)-measures.

Double track When researching heterogeneity, it is quite obvious that one has to look at double track sections. The timetable for single-track lines is mostly dictated by distances between passing points. In the case of four tracks, which means two tracks per direction, trains with different speeds are already separated, and each track has its own speed: one track for slow traffic, one for fast traffic. The interesting part is where all trains for one direction run on one track: double track lines. Notice that many of Europe's main lines are double track lines indeed.

Experiments For the experiments in this chapter we have chosen for the option of shifting stops from the short distance services to the long distance services in order to equalize the number of stops for more homogeneous timetables.

The cyclic timetables are developed with the timetabling tool DONS, which is described in Section 3.1. For the comparison of the timetables, simulation of railway traffic has been used. The simulations reported on are performed with SIMONE. This simulation tool was portrayed in Section 3.3.

5.3 A Theoretical Case

The first case that we are looking at is a theoretical case. Both the simple network and the timetable are artificial and have been developed especially for the heterogeneity comparison.

5.3.1 Case Description

The network consists of two intersecting double-track lines of 192 kilometer each, which intersect at a Central Station (CS). This creates four identical branches of 96km: northwest (NW), northeast (NE), southwest (SW), and southeast (SE). The layout of the network is shown in Figure 5.3(a), where the lines do not represent tracks but train lines. The four branches are equal and have three intermediate large stations, where all trains stop. These stations are represented by the rectangles in Figures 5.3(b) and 5.3(c). Each branch also has ten smaller stations, where half of the trains stop. These small stations are closer to each other around CS and around the endpoints. One can consider these areas as denser populated. The distances were chosen such as to resemble the average station distance in the Netherlands. Only the stations where a train line stops are shown: in Figure 5.3(b) the heterogeneous situation, and in Figure 5.3(c) the homogeneous situation. Note that the number of trains and the number of stops per station are equal for these two situations.

The intersecting lines at Central Station have free-level crossings only. This means that only trains going to or coming from the same direction can interfere with each other. Still, delays can be transferred throughout the network because of the long distance trains, which alternate in destination. Figure 5.3(a) shows which direct connections exist in the experiments: each line represents two trains per direction per

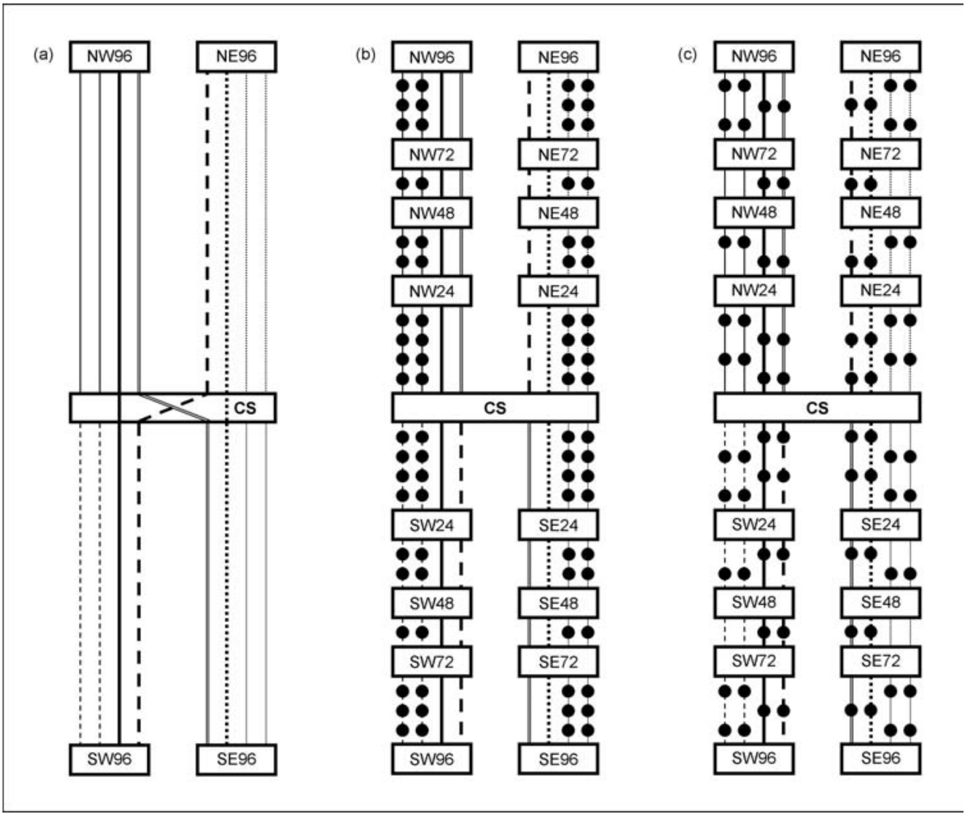


Figure 5.3: (a) shows the different train connections in the theoretical network, with the heterogeneous dwelling pattern in (b) and the homogeneous dwelling pattern in (c).

hour. There are four short distance trains on each branch from CS to the endpoint. Additionally, there are four long distance trains per hour on each branch, but they alternate in destination: there are two trains per hour from NW96 to SW96, two from NW96 to SE96, two from NE96 to SE96, and two from NE96 to SW96.

Although generally different types of rolling stock are deployed for short distance trains and long distance trains in real life, only one type of rolling stock is used for simplicity in this case.

Heterogeneous situation In the heterogeneous situation, the short distance trains stop at all stations and the long distance trains only dwell at the large stations. This dwelling pattern is shown in Figure 5.3(b). All trains are nicely spread over the hour, which means that a short distance train leaves from every station in the system exactly every 15 minutes in both directions. At the large stations, one can also catch a long distance train exactly every 15 minutes, where the individual long distance lines (for example NW96-SW96), run exactly every 30 minutes. A time-distance diagram created by DONS is given in Figure 5.4. This timetable includes exactly 7% running time supplement on each trip.

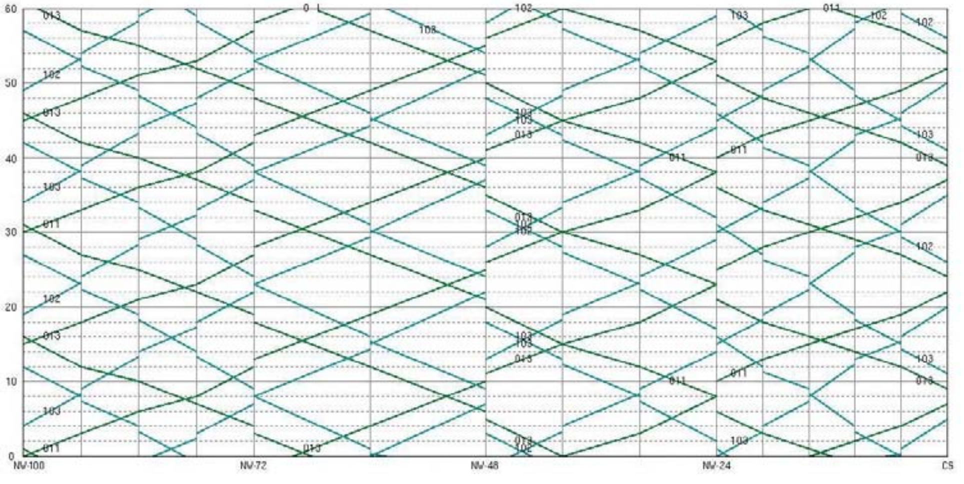


Figure 5.4: Time-distance diagram of the heterogeneous situation for the branch NW96-CS: there is an apparent difference between the long distance trains (the flatter lines) and the short distance trains (the steeper lines). Vertical jumps in the lines depict stops at stations. Due to acceleration, deceleration and roundings, running times on tracks of equal length may differ. The other three branches have identical time-space diagrams.

The SSHR between CS and one of the endpoints is 5.33 for the heterogeneous case. The SAHR can only be defined for single locations, leading to multiple values. At CS the SAHR is 2.31 from all directions. Coming from CS, the SAHR is also 2.31 at NW48, NE48, SW48 and SE48. Coming from the endpoints, the SAHR is 1.67 at NW48, NE48, SW48 and SE48. Finally, the SAHR is 1.36 at the endpoints.

It should be noted that in heterogeneous situations, overtaking may have to take place, depending on the frequency, the difference in numbers of stops and the time loss per additional stop. In the presented case, overtaking is necessary indeed: the short distance trains are overtaken by the long distance trains in NW48, NE48, SW48 and SE48. Furthermore, the dwell time of the long distance trains is extended at NW24, NE24, SW24 and SE24 to decrease the travel time differences. Otherwise no feasible timetable would exist for these trains, these dwellings, and this train order.

Homogeneous situation In the homogeneous situation, the same number of lines, the same number of stops per station, and the same line-endpoint connections are applied as in the heterogeneous situation. Also the same type of rolling stock is used. However, a more homogeneous situation is created by decreasing the number of stops of the short distance services and increasing the number of stops of the long distance services until both are as equal as possible. The newly created services are shown in Figure 5.3(c).

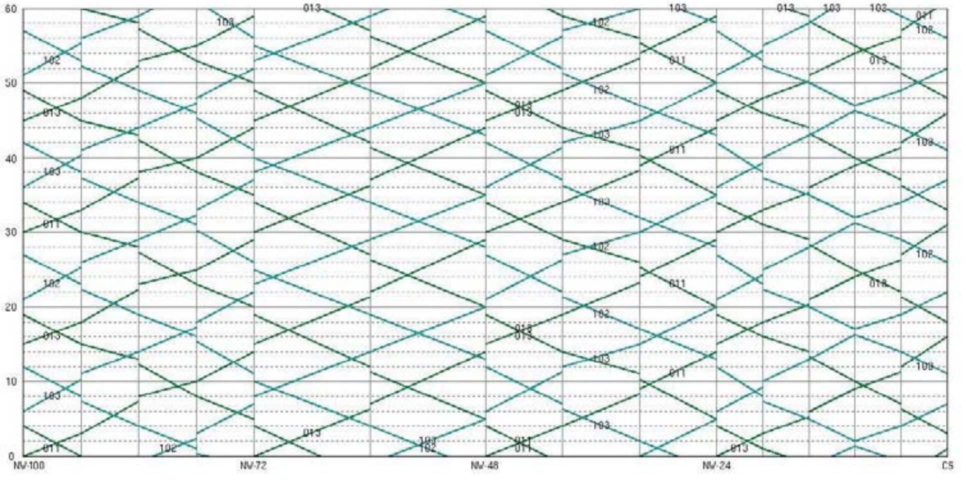


Figure 5.5: Time-distance diagram of the homogeneous situation: the different trains have almost parallel time-distance diagrams, which leads to a larger minimal headway between trains.

Per branch, starting from the endpoints, the long distance lines dwell at the 2nd, 4th, 6th, 8th, and 10th small station, whereas the short distance lines (NW96-CS and so on) dwell at the 1st, 3rd, 5th, 7th, and 9th small station. Again, every small station is served four times an hour, exactly every 15 minutes, and each intercity station is served eight times per hour, as is shown in Figure 5.5. Also this timetable includes exactly 7% running time supplement on each trip.

The homogeneous situation reduces the headway measures significantly. The SSHR goes down from 5.33 to 1.24. At CS the SAHR decreases from 2.31 to 1.11. At NW48, NE48, SW48 and SE48, we see the SAHR dropping from 2.31 to 1.07 (from CS), and from 1.67 to 1.11 (from the endpoints). The SAHR at the endpoints decreases from 1.36 to 1.07.

5.3.2 Experimental Design

Sixteen experiments with different disturbance distributions were carried out. Disturbances were generated randomly by SIMONE, following the specifications given in Table 5.1. Dwell times and running times are disturbed with a certain probability. These disturbances are exponentially distributed with a given average. Earlier research (Goverde et al., 2001) shows that exponential distributions fir well to the real dwell disturbances and late arrivals.

Experiment	Dwell Time Disturbance all stations: probability of disturbance	Dwell Time Disturbance all stations: average size in minutes	Dwell Time Disturbance large stations: probability of disturbance	Dwell Time Disturbance large stations: average size in minutes	Absolute Running Time Disturbance: probability of disturbance	Absolute Running Time Disturbance: average size in minutes	Total Disturbances in minutes per hour
1	5%	1					30.8
2	5%	1	5%	2			61.7
3	5%	1	10%	2			92.3
4	5%	1	15%	2			123.0
5	5%	1	20%	2			152.9
6	5%	1	25%	0.5			69.5
7	5%	1	25%	1			107.6
8	5%	1	25%	1.5			146.2
9	5%	1	25%	2			183.4
10			25%	1			76.3
11			25%	1	20%	0.075	128.1
12			25%	1	40%	0.075	179.5
13			25%	1	60%	0.075	232.2
14			25%	1	10%	0.15	128.5
15			25%	1	20%	0.15	179.3
16			25%	1	30%	0.15	232.1

Table 5.1: Experimental design for the theoretical case

All simulation experiments consist of fifty runs of 1320 minutes, including 120 minutes of warm-up time. This leaves exactly fifty times twenty hours of simulation time for which statistics are collected. Twenty hours is close to one day of train services in the Netherlands, where night services are almost non-existing. Fifty runs leads to a satisfactory reliability of the simulation results, where the average delays have standard deviations up to 3% for the heterogeneous situation and up to 7% for the homogeneous situation (except for experiment 1, where small probabilities for disturbances cause a higher variability of the results).

The first experiments have a combination of dwell time disturbances at all stations and dwell time disturbances at large stations. This leads to two cumulative disturbances for the large stations. Experiments 1 to 5 have an increasing probability for disturbances at large stations; experiments 1 and 6 to 9 have an increasing average size of dwell time disturbances at large stations.

The following experiments have a combination of dwell time disturbances at large stations and running time disturbances. Experiments 11 to 13 have the same total number of disturbance minutes as experiments 14 to 16, but the latter experiments have fewer, though larger, disturbances.

Experiment	Average Arrival Delay (per train measurement in minutes)			3-minute Unpunctuality (% of trains delayed)			Incurred Secondary Delays (in minutes per hour)		
	heterogeneous situation	homogeneous situation	improvement (relative in %)	heterogeneous situation	homogeneous situation	improvement (relative in %)	heterogeneous situation	homogeneous situation	improvement (relative in %)
1	0.38	0.09	76.5	2.1	0.5	75.0	36.9	0.5	98.7
2	0.97	0.27	72.4	6.9	2.5	64.2	88.2	4.4	95.0
3	1.57	0.46	70.7	11.7	4.8	59.3	136.0	9.1	93.3
4	2.16	0.65	69.9	16.1	6.9	57.0	177.5	13.4	92.4
5	2.70	0.88	67.5	20.1	9.7	51.7	212.5	18.3	91.4
6	0.64	0.15	76.8	3.8	0.7	82.5	69.9	0.6	99.2
7	1.38	0.33	76.3	9.6	2.3	76.0	129.0	2.1	98.4
8	2.38	0.65	72.5	17.3	6.6	61.8	195.8	9.8	95.0
9	3.21	1.08	66.4	23.6	12.2	48.3	243.1	22.5	90.7
10	0.92	0.21	77.0	5.9	1.2	79.1	83.9	1.1	98.7
11	1.28	0.32	75.3	8.3	1.9	77.0	117.3	1.5	98.8
12	1.73	0.46	73.4	11.4	3.0	73.8	152.6	2.0	98.7
13	2.28	0.66	70.9	15.4	4.7	69.8	188.8	3.0	98.4
14	1.53	0.38	75.0	10.4	2.5	75.8	135.9	2.1	98.5
15	2.20	0.60	71.0	15.4	4.7	69.8	184.4	3.7	98.0
16	2.99	0.88	70.6	21.3	8.0	62.4	234.9	6.0	97.4

Table 5.2: Simulation results for the theoretical case. The presented average delays have standard deviations between 0.26% and 2.65% in the heterogeneous case, and between 0.76% and 7.00% in the homogeneous case, with the exception of experiment 1 (6.21% and 18.38% respectively).

5.3.3 Results

- Table 5.2 shows that, going from the heterogeneous to the homogeneous situation, the average delay decreases by over 65% in all experiments. Although the decrease in unpunctuality is varying quite a bit, the average delay always decreases by 66 to 77 percent in the model.

- The last three columns of Table 5.2 show where the differences originate from: the secondary delays. Especially in experiments with only few and small disturbances, there are hardly any secondary delays in the homogeneous situation. A simple explanation for this is the fact that secondary delays will only emerge if there are 'large' disturbances. The smallest planned headway between trains in the homogeneous situation is six minutes, and the minimally required headway is only two minutes. This allows for a disturbance of at least four minutes for a certain train before it causes secondary delays to subsequent trains. In the heterogeneous situation, the slightest disturbance causes secondary delays already, because trains are scheduled at minimal headway distance.

- When the average disturbances are smaller, the relative improvement is higher. This is because the homogeneous situation is relatively immune for small disturbances, as was explained in the paragraph above. When the average disturbances increase, secondary delays also occur in the homogeneous situation, reducing the relative difference with the heterogeneous situation.

- Additionally, few large disturbances are more harming to the punctuality than many small disturbances with the same total number of minutes of disturbances. Compare for example experiments 11, 12 and 13 with experiments 14, 15 and 16. The total disturbances in minutes are equal, but the average size of a running time disturbance is twice as large in the latter experiments. This leads to an average delay increase of 20% to 30%.

  Two explanations may be viable here. First, the explanation above is valid again: where a small disturbance is too small to disturb a second train, this is valid for two small delays on two different trains as well. However, one large disturbance may have a negative effect on other trains. Secondly, two separate disturbances are recovered by two separate running time supplements. Suppose in situation 1, that one train, running from A via B, C and D to E, is disturbed 2 minutes at station A and between every pair of stations half a minute of delay can be recovered. Then there will be a 1.5 minute delay at station B, one minute at C, half a minute at D, and the train will arrive on time at station E. The average arrival delay is $(1.5 + 1 + 0.5 + 0)/4 = 0.75$ minutes. A second train, also running from A to E, is not disturbed at all. The average arrival delay of these two trains is 0.375 minutes. Now in situation 2, let us take two trains for the same line, which both have a one-minute disturbance at A. They will both be 30 seconds delayed at station B, but they will arrive on time at stations C, D and E. This results in an average arrival delay of only $2(0.5 + 0 + 0 + 0)/8 = 0.125$ minute.

- The dwell time disturbances at all stations have the largest impact on trains with many stops. This means that the expected travel time increase is largest for the short distance trains in the heterogeneous situation. Therefore, the expected travel time differences are even larger than the planned travel time differences.

Note that the simulations have been carried out without any other dispatching rule than first-come-first-serve. Because we are also only evaluating the impact of small disturbances, the impact of dispatching is probably relatively small. Still, the consequences of other dispatching rules are unknown.

Causes for delays in the heterogeneous situation Finding out where the delays come from can provide interesting information. Therefore, the average delays over the course of the train services are examined here. Figure 5.6 shows the average delay of the four branches, separated for trains running towards CS (indicated by 'a'), and running from CS (indicated by 'b'). This is the graph for experiment 3, but it is typical for all experiments in the heterogeneous case.

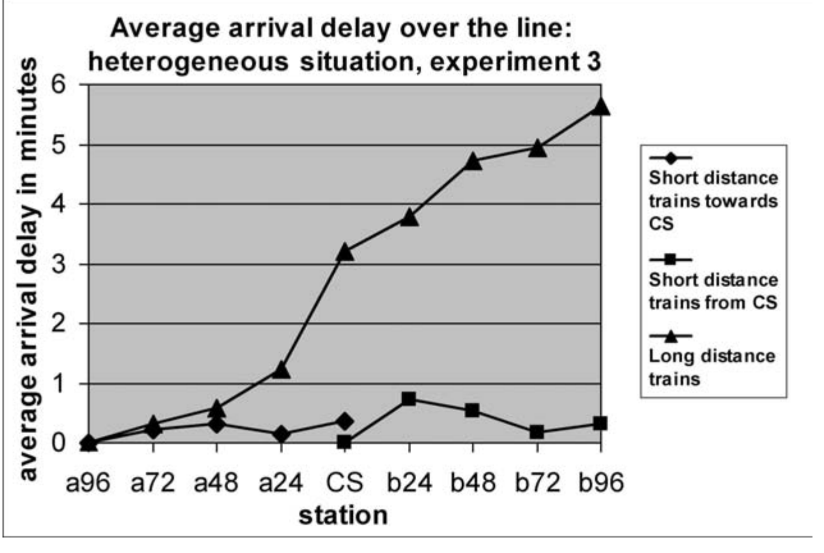


Figure 5.6: Average delay over the course of the lines. For example, a72 shows the average arrival delay at NW72, NE72, SW72 and SE72 for trains towards CS. At b48 the average delay for arrivals at NW48, NE48, SW48 and SE48 is given for trains going towards the endpoints.

- The first thing to notice is the fact that the short distance trains are hardly delayed at all, while long distance trains have considerable delays. This is because a long distance train can be caught behind a delayed short distance train. In that case, the long distance train can incur a large secondary delay, because it cannot run any faster than the short distance train up to the next station where overtaking is possible. Short distance trains can never be caught behind slower trains. This is exactly why the SAHR was developed next to the SSHR. The largest short distance train delays are found at the first station after starting from CS. This is explained by the delays of the long distance trains in CS, which hamper the departure of the short distance trains starting from CS.

- The average delay of the long distance trains increases with distance traveled, but not linearly. The largest delay increases can be found just before CS and b48. This is exactly where the heterogeneity causes the headways to be smallest (see Figure 5.4) and the intercity trains catch up with the short distance trains. The two other locations where the intercity trains catch up with the short distance trains, a48 and b96, seem to cause less problems. This is due to the lower heterogeneity, and consequently the larger headways, in the peripheral areas.

5.4 A Practical Case

Besides the theoretical case presented above, a practical case has been worked out as well. We have compared a real life heterogeneous timetable with a more homogeneous timetable for a busy line in the Netherlands. Some details of the real timetable have been adjusted slightly for the simulations (NS, 2002).

5.4.1 Case Description

The case which is elaborated here consists of the lines from The Hague Central (Gvc) and Rotterdam Central (Rtd) to Utrecht Central (Ut), which merge at Moordrecht Junction (Mda). These lines are represented by the bold lines in Figure 5.7. For the case, this part of the network has double track everywhere, except for the section between Moordrecht Junction and Gouda Goverwelle (Gdg), which has four tracks. Moordrecht Junction is a non-level crossing. The distance between The Hague and Utrecht is 61 km, the distance between Rotterdam and Utrecht is 56 km.

In Figure 5.8, all lines have a cycle time of 30 minutes, which leads to, for example, eight trains per hour between Rotterdam Central and Moordrecht Junction, and twelve trains per hour between Woerden (Wd) and Utrecht Central.

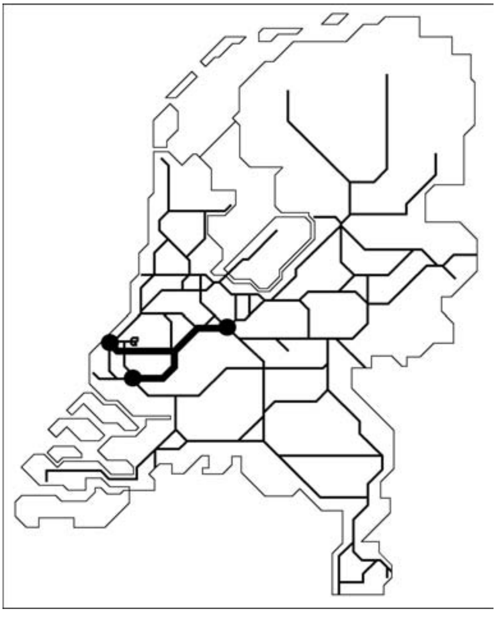


Figure 5.7: The railway network served by NS. The bold lines represent the tracks between The Hague and Utrecht and between Rotterdam and Utrecht.

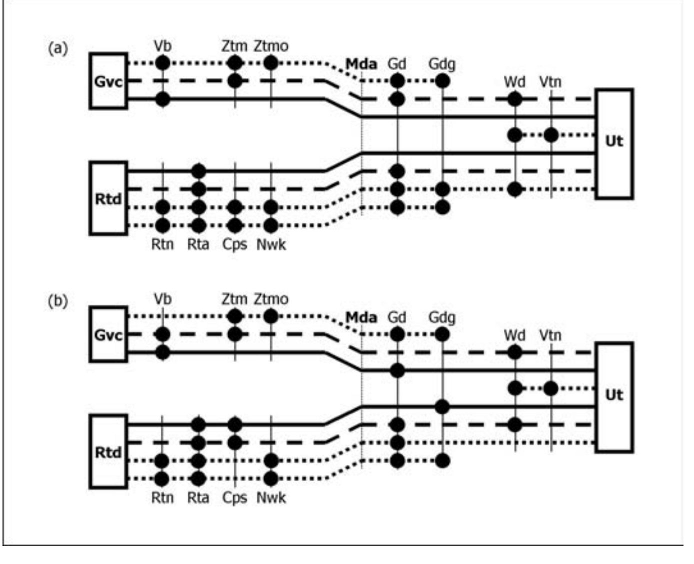


Figure 5.8: Train lines and dwelling patterns of the heterogeneous situation (a) and the homogeneous situation (b) of the practical case. Except for the junction Mda, all abbreviations indicate a station.

Heterogeneous situation With some adjustments, the 2003 rush-hour timetable has been taken for the heterogeneous situation (NS, 2002). Cargo trains are skipped, resulting in a three-train-system: long distance trains, interregional trains and short distance trains, represented by solid (——), broken (– – –), and dotted lines (⋯⋯), respectively, in Figure 5.8.

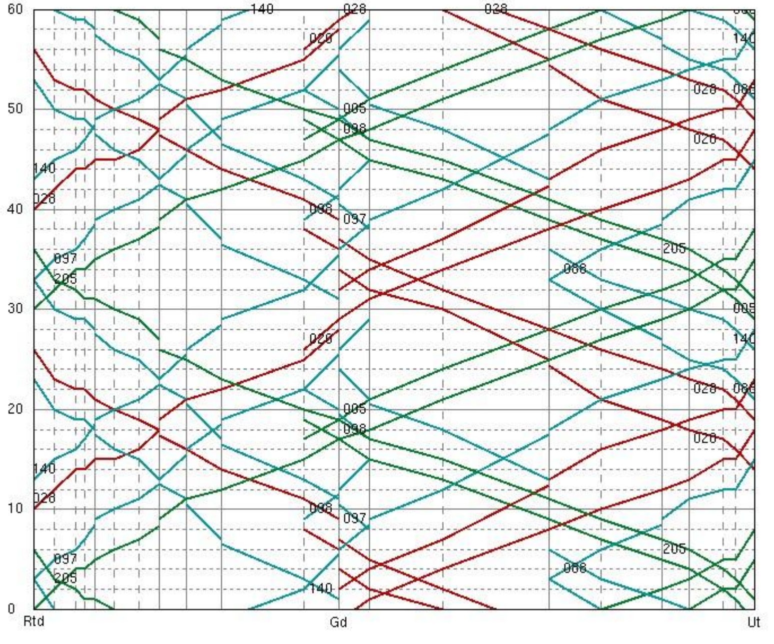


Figure 5.9: Time-distance diagram for the heterogeneous situation on the Rotterdam-Utrecht branch.

Every 30 minutes there is one short distance train from The Hague to Gouda Goverwelle, there is one interregional train from The Hague to Utrecht, and there is one long distance train from The Hague to Utrecht. Starting from Rotterdam, there is one short distance train running to Gouda Goverwelle, a second short distance train running to Utrecht (not dwelling in Vleuten (Vtn)), one interregional train to Utrecht, and one long distance train to Utrecht. Additionally there is a short distance train from Woerden to Utrecht. This adds up to 16 trains per hour per direction. The time-distance diagram for Rotterdam–Utrecht is shown in Figure 5.9.

Unlike the theoretical case, the lines in this case are operated by different rolling stock types. These rolling stock types have their own specific characteristics concerning acceleration and top speed. These are, according to the real life situation, matched with the service provided.

To be able to compare the two timetables on their heterogeneity, both timetables were adjusted such that each trip includes exactly 7% running time supplement.

Homogeneous situation Like in the theoretical case, the heterogeneous situation is homogenized by decreasing the number of stops of the short distance services and compensating those by additional stops of the faster services. In the end, the total number of stops per station is equal in both situations. The final dwell pattern is shown in Figure 5.8(b), the time-distance diagram is shown in Figure 5.10.

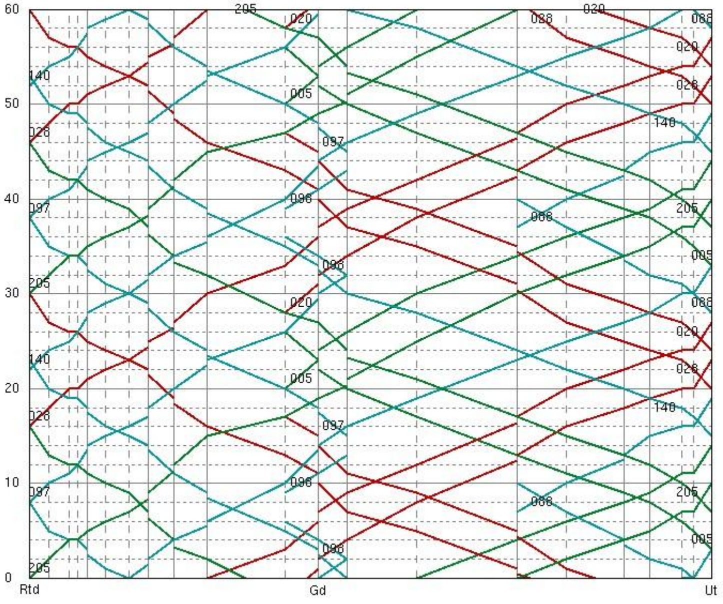


Figure 5.10: Time-distance diagram for the homogeneous situation on the Rotterdam-Utrecht branch.

Because of the homogenization, there is no clear distinction anymore between slower and faster services. Therefore, the necessity for different types of rolling stock has gone. However, for a fair comparison, the same rolling stock has been used for both situations.

The SSHR and SAHR for the practical case are given in Table 5.3 below. On all sections, both the SSHR and the SAHR are lower for the homogeneous situation.

SSHR heterogeneous situation	SSHR homogeneous situation	trajectory	SAHR heterogeneous situation	SAHR homogeneous situation
0.76	0.67	Gvc-Gd	0.64	0.62
3.18	1.47	Rtd-Gd	1.36	1.17
3.47	2.97	Gd-Ut	2.70	2.52
5.52	3.02	Ut-Gd	2.55	1.77
1.06	0.69	Gd-Gvc	0.67	0.61
2.83	1.19	Gd-Rtd	1.74	1.07

Table 5.3: The SSHR and SAHR for the different tracks between The Hague, Rotterdam and Utrecht.

5.4.2 Experimental Design

Again, the simulation experiments consist of fifty runs of 1320 minutes, including 120 minutes of warm-up time. The disturbance distributions, all exponential again, are given in Table 5.4.

Experiment	Dwell Time Disturbance all stations: probability of disturbance	Dwell Time Disturbance all stations: average size in minutes	Absolute Running Time Disturbance: probability of disturbance	Absolute Running Time Disturbance: average size in minutes	Relative Running Time Disturbance: probability of disturbance	Relative Running Time Disturbance: average size % running time	Total Disturbances: in minutes per hour
1	100%	0.6					95.9
2	100%	0.8					128.0
3	10%	1	5%	1			38.7
4	20%	1	10%	1			78.2
5	50%	0.5	20%	0.5			86.2
6	75%	0.5	30%	0.5			129.2
7					30%	40%	112.8
8					60%	20%	114.0
9	20%	1			30%	20%	89.3
10	50%	0.5			30%	20%	96.9

Table 5.4: Experimental design for the practical case

The first two experiments have only dwell time disturbances at all stations. The next four experiments have a combination of dwell time disturbances and absolute running time disturbances. Absolute running time disturbances are independent of the running time and have the averages given in Table 5.4. The disturbances of experiments 5 and 6 have a larger probability of occurring, but are, on average, smaller in size than those of experiments 3 and 4.

Experiments 7 and 8 have relative running time disturbances. These relative disturbances depend on the planned running time of the train on the track where it is disturbed. The average is equal to a certain percentage of the running time. Although the total disturbances are equal, experiment 8 has more but smaller disturbances than experiment 7. Experiments 9 and 10 have both dwell time disturbances and relative running time disturbances.

The results from the simulations are given in Table 5.5 and are comparable with those of the theoretical case. The main distinction is that the decrease in delays between the heterogeneous and homogeneous situation is smaller. This is easily explained by the smaller differences in heterogeneity in the practical case.

Experiment	Average Arrival Delay (per train measurement in minutes)			3-minute Unpunctuality (% of trains delayed)			Incurred Secondary Delays (in minutes per hour)		
	heterogeneous situation	homogeneous situation	improvement (relative in %)	heterogeneous situation	homogeneous situation	improvement (relative in %)	heterogeneous situation	homogeneous situation	improvement (relative in %)
1	1.08	0.72	33.4	8.8	3.1	64.6	30.0	3.2	89.3
2	1.85	1.22	34.0	21.1	10.4	50.8	50.3	8.4	83.4
3	0.60	0.40	33.7	5.0	3.0	39.9	20.9	4.2	80.2
4	1.40	0.92	34.5	15.0	9.0	39.9	41.3	10.8	74.0
5	1.13	0.75	33.6	8.7	3.9	55.4	28.8	3.9	86.6
6	2.02	1.35	33.2	23.1	12.0	48.2	47.5	8.7	81.7
7	2.21	1.55	29.9	25.9	17.6	32.1	60.2	22.9	62.0
8	1.71	1.22	28.6	18.3	10.8	41.1	38.7	10.4	73.1
9	1.30	0.89	31.4	12.4	7.5	39.6	35.0	8.7	75.2
10	1.28	0.88	31.0	11.4	5.9	48.0	32.1	6.3	80.4

Table 5.5: Simulation results for the practical case. The presented average delays have standard deviations between 0.3% and 2.4% in the heterogeneous case, and between 0.3% and 3.3% in the homogeneous case.

5.4.3 SSHR, SAHR and the Results

In Figures 5.11 and 5.12, the SSHR and SAHR improvements are compared with the average delay reduction. For this comparison, the network is divided into six sections: The Hague-Gouda, Rotterdam-Gouda and Gouda-Utrecht and vice versa.

The average delay for each section, as given in the figures, is the average delay of all trains at the endpoint of the section minus the average delay of these trains at the start of the section.

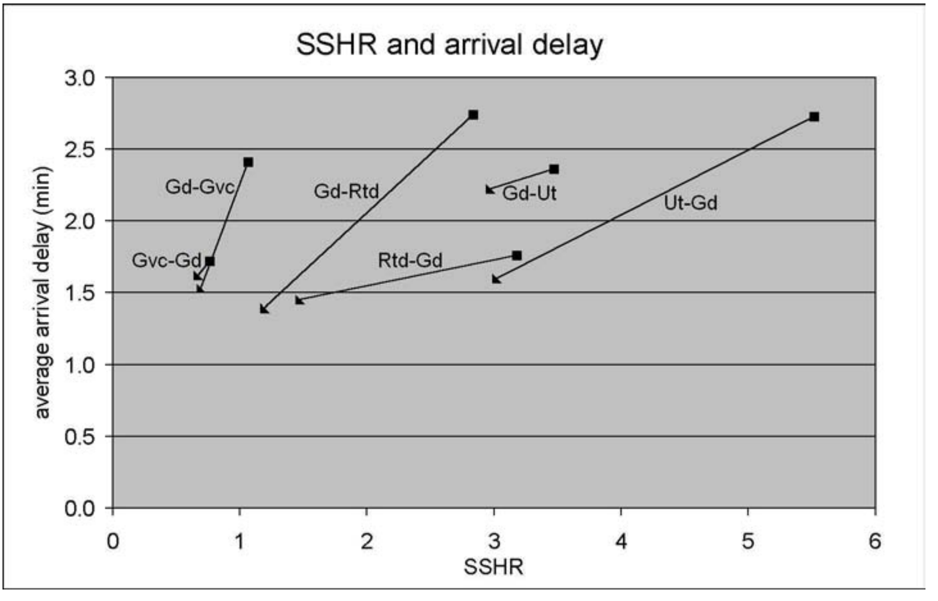


Figure 5.11: The relation between the decrease in SSHR and the reduction of the average delay.

The upper right of each line segment in Figure 5.11, represented by a square, is the SSHR and average delay in the heterogeneous situation. The other endpoint, represented by an arrowhead, is the result of the homogenization. Figure 5.12 shows the same comparison for the SAHR. The graphs represent the results of experiment 7. The figures for the other experiments are quite similar.

- Figures 5.11 and 5.12 show that homogenization of the timetable leads to both a reduction in the SSHR and SAHR measures, and a reduction in average delay for all sections.
- The decrements in SSHR are quite different for the different sections: The Hague-Gouda only shows a small difference, while Utrecht-Gouda shows a large decrease. The same is true for the SAHR.
- The relative reductions in SSHR are not equal, nor almost equal, to those of the SAHR. This means that SSHR and SAHR are two quite distinct measures. See for example Rotterdam-Gouda.

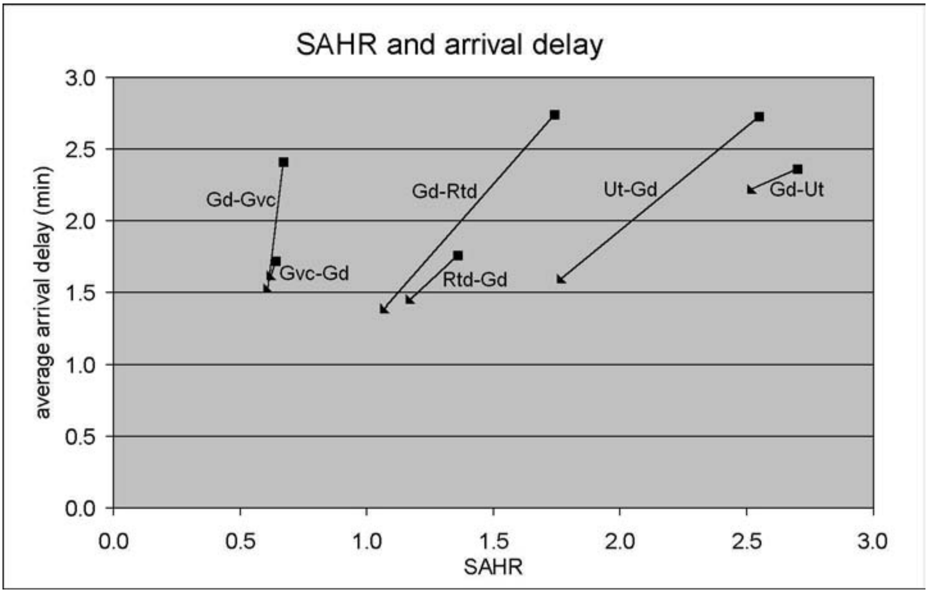


Figure 5.12: The relation between the decrease in SAHR and the reduction of the average delay.

- In general, a larger decrease in SSHR leads to a larger reduction in delays. Still, the line segments for Rotterdam-Gouda and Gouda-Utrecht are rather flat. This means that the delays do not decrease as much as the SSHR might indicate. Therefore the SSHR can be used as an indication in what direction the reliability goes, but it is not an absolute measure. For the starting sections, the departure reliability from the first stations is very high. This means that small departure headways hardly have a negative influence on the reliability. The reduction of the SSHR on Rotterdam-Gouda is mainly based on a fairer distribution of departures; arrivals are hardly affected by the homogenization. This may explain the relative small reliability improvement of this section.

- Also, a larger decrease in the SAHR leads to a larger reduction in delays. Still, sections with a relatively large delay reduction (a steep line segment) can be observed (e.g. Gouda-The Hague), as well as sections with a relatively small reduction (e.g. Gouda-Utrecht). Due to the large reduction of the delays on the section Utrecht-Gouda, predicted by the SAHR, there is also a large reduction in departure delays for the section Gouda-The Hague. A better starting reliability implies fewer secondary delays, which explains the large delay reduction on Gouda-The Hague, and to a lesser extent, on Gouda-Rotterdam.

- Both Figure 5.11 and Figure 5.12 have, at least on average, a lower bound for the average delay on each section. This is due to the amount of disturbances, which exceeds the amount of running time supplements. The running time supplements equal 7% of the running times, whereas the disturbances average 12% of the running times. This means that running times are, on average, extended by 5%. Because all considered sections have running times of around 20 minutes, this amounts to approximately 1 minute as the lower bound for the average delay.

5.5 Discussion

The measures: SSHR and SAHR The measures SSHR and SAHR are both able to predict reliability changes. The SSHR is applicable to track sections between stations, whereas the SAHR can be used for a station, or for all arrivals from a certain track at a station. Although a large decrease in the SSHR or in the SAHR leads to a large decrease of delays in most cases in the model, it is hard to predict the exact size of the delay reduction. Using a weighted average of the two measures may be advantageous, because it takes heterogeneity into account, and it weighs the arrival headways more heavily than the departure delays.

The measures ASSBR($\overline{\mathrm{h}^{\mathrm{min}}}$) and ASABR($\overline{\mathrm{h}^{\mathrm{min}}}$) can be used to replace the SSHR and SAHR, respectively, when there are (large) differences in technically minimal headways.

Equalizing headways Minimizing the SSHR or the SAHR implies equalizing the headways. Although a reduction in these measures indicates a reduction in delays, minimizing the sum of SSHRs or SAHRs over the network is not necessarily optimal. This can for example be seen from the SSHR in the practical case, where a large reduction on one section (Rotterdam-Gouda) has much less influence on the reliability than a small reduction of the SSHR on another track (Gouda-The Hague).

Utilize and Build "Utilize and Build" is the vision and the intended direction of the combined Dutch railway sector for the future, up to 2020. Experts from the Ministry of Transport, the railway infrastructure manager ProRail, the passenger operator NS, and several cargo operators participate in this project (NS et al., 2003). The main problem is how to facilitate the ever expanding railway traffic on the limited infrastructure. The starting point of the project is to better utilize the existing infrastructure, which is facilitated by small but smart infrastructure investments. Homogenization of the railway system is one of the basic elements of "Utilize and

Build". Although the main focus is on a limited homogenization where intercity trains stop at a few more stations of medium large size, full homogenization is also discussed and not excluded a priori as a solution.

Other consequences of homogenization Although reliability will increase when train services are homogenized, there are several other important characteristics to be considered both for passengers and operators. Homogenization can have its influence on many of those characteristics.

- Travel time for passengers is an important determinant of service quality in case of homogenization. The planned travel time may decrease for some passengers, but increase for others. The number of passenger transfers and the transfer times may also change. This requires a further mobility analysis, which falls outside the scope of this thesis.
- Infrastructural needs can possibly change due to other train lengths, but also due to other locations for overtakings, and due to another way of coordinating trains at large transfer stations.
- When the timetable is homogenized, the rolling stock can be standardized as well. The total required number of rolling stock units can also change.
- Homogenization by one large operator may lead to additional time-slots in the timetable, which might be assigned to other operators. Evidently this would, in the end, lead to an increased SSHR and SAHR. Therefore, network wide cooperation is necessary for a beneficial introduction of a homogenized timetable.

5.6 Conclusions

In this chapter two timetable characteristics, the Sum of Shortest Headway Reciprocals (SSHR) and the Sum of Arrival Headway Reciprocals (SAHR), were introduced. These measures can be used for evaluating the heterogeneity of the timetable and for the prediction of the reliability. The SSHR can be applied to a whole railway section and has the desirable property that it decreases both when trains are spread better over the hour and when railway traffic is more homogeneous. The SAHR also has the property of decreasing when the trains are spread better over the hour. It lacks a direct link to the heterogeneity, but takes it into account implicitly. In case of differences in technically minimal headways, the SSHR can be replaced by the ASSBR($\overline{\mathrm{h}^{\mathrm{min}}}$), and the ASABR($\overline{\mathrm{h}^{\mathrm{min}}}$) can be used instead of the SAHR.

The presented cases show a large reliability increase for homogenized services with low values for the SSHR and SAHR. In other words, when the SSHR and SAHR show large decreases, then there are usually also large decreases in delay propagation. Therefore, a relatively simple rule of thumb for timetable design is to minimize the SSHR and the SAHR, at least per section. This may improve the reliability of the offered services.

Although homogenization may lead to a sizable increase in punctuality of the offered railway services, homogenization may also affect other features of the railway product, both for passengers and for operators and infrastructure managers. When homogenizing train services, these other consequences should also be considered. This is a subject for further research. The relationship between the consequences for the different operators and the infrastructure managers also stresses the importance of cooperation between these parties.

Chapter 6

Stochastic Timetable Optimization

When a cyclic timetable is constructed, it can be evaluated afterwards in several ways, for example with max-plus algebra or using simulation. However, this kind of evaluation does not tell anything about the optimality of the timetable and it certainly does not provide alternatives.

Therefore, a model which creates an optimal timetable itself is preferred. In this chapter, we present a delay minimizing timetabling model, based on a stochastic optimization model with fixed recourse. For a predetermined set of primary delays, this model creates an *optimal* timetable with respect to the average delay. Besides primary disturbances, secondary delays are also taken into account. Furthermore, other timetabling characteristics, such as travel times, can also be considered.

First, it has to be noted that finding an optimal timetable implies that the timetable still has to be constructed. Therefore a timetabling model, such as the PESP-model in Section 3.1, is required in the model.

Secondly, the delay propagation has to be evaluated in the model to be able to minimize the average delay. For known timetables, several options are available, such as simulation (see Section 3.3) and max-plus algebra (see Section 3.2). Simulation has the advantage that a wide range of stochastic processes can be modeled and evaluated. Unfortunately, the structure of a regular simulation model is hardly comparable to that of a timetabling model. This implies difficulties for integrating the two. The structure of max-plus railway models is, rewritten in classical algebra, quite similar to that of cyclic timetabling models with a fixed order of events. However, the use of stochastic disturbances in max-plus algebra is hard.

The model that is described in this chapter is a stochastic optimization model (see e.g. Klein Haneveld and Van der Vlerk (2004), Birge and Louveaux (1997)), integrating a timetabling model based on PESP, and a reliability evaluation based on simulation with a fixed order of events.

The resulting timetable optimization model is described in this chapter. It is both capable of constructing a timetable and of calculating the delay propagation. A number of consecutive realizations is performed by repeating the periodic timetable under construction several times. These realizations are perturbed by exogenous random disturbances, and the resulting arrival delays are measured. Moreover, the timetable is optimized with respect to the average delay.

The freedom in the timetable which has to be optimized consists of several dimensions. Important are the planned running times and the resulting running time supplements. These supplements can absorb incurred delays. The same can be said about the supplements in dwell times and the slack in passenger transfers and layover times. Also, the buffer times between trains can be adjusted by the model so that delay propagation is minimal. Furthermore, the train order on the tracks is an important determinant of the delay propagation. A special variant of the model can also decide on the dwell locations of the lines to improve the reliability.

Section 6.1 discusses the relation between our delay minimizing model and recourse models. In Section 6.2 the model is introduced using the supplement allocation problem of Chapter 4 as an example. For a limited situation, the convergence of this model is proven in Section 6.3. In Section 6.4 it is shortly described how the stochastic influences by the disturbances can be reduced by using smart sampling methods. This can speed up the convergence of the model. Section 6.5 is the core of this chapter. It describes the model in a more general form, and extends it so that railway timetables can be modeled and optimized with respect to the average delay. The Haarlem – Maastricht case in this section shows that, within the model setting, delay reductions of over 30% are possible.

The later parts of this chapter discuss extensions to the stochastic model. Section 6.6 discusses some alternative objectives. In Section 6.7 binary variables are introduced to relax some of the assumptions made in Section 6.5. This complicates the optimization, but allows us to model almost all standard timetable situations. Furthermore, the model is extended to model variable dwell patterns in Section 6.8, which enables the model to homogenize the railway traffic. We have already seen in Chapter 5 that this may lead to a sizable delay reduction. In the last section the model and its results are discussed and conclusions are drawn.

6.1 Relation with Recourse Models

Recourse models can be used to model decisions that have to be taken under uncertainty. A recourse model is a stochastic programming model that can be seen as a two-stage model (Klein Haneveld and Van der Vlerk, 2004). In the first stage, decisions have to be made before stochastic events occur. These stochastic events influence the success of the decisions made earlier in the model. These stochastic influences and the resulting outcomes are the second stage of the model. The decision maker may have to pay a high price in the second stage if he was too optimistic about the outcome of the stochastic events. However, too conservative assumptions about the outcome of the stochastic events may lead to high costs in the first stage. Recourse models help the decision maker to find an optimum between these two extremes.

In many practical situations, a repetitive interaction between decisions and random events takes place. This can be modeled by a multistage recourse model. In our case, we use a two-stage recourse model. In the first stage we develop an *optimal* timetable. In the stochastic second stage we pay for the delays resulting from this timetable. The cyclicity of the timetable restricts the problem to a two-stage problem: one does not want to adjust the schedule over and over again, depending on past disturbances. Therefore, there is no interaction anymore between decision maker and planned timetable after the first stage. The second stage is used for the evaluation of the timetable under construction.

To be a little more specific, our timetabling model is a *stochastic programming model with recourse*. The general form of this model can, for example, be found in Klein Haneveld and Van der Vlerk (2004) or Birge and Louveaux (1997). Note that the notation of the parameters and variables in this section are not in line with the remainder of this thesis, but with most literature on stochastic programming.

Model notation The first stage decision variables are denoted by x, and c^T is the cost vector associated with the first stage decision variables. The linear system $Ax \geq b$ has to be satisfied in the first stage. In practice, the vector x has to be determined before the stochastic outcomes of the model are known. These stochastic outcomes represent external influences on the model environment. These stochastic outcomes can influence the relations between the variables (represented in T and W), as well as the parameter values of the right-hand-side of the restrictions: h. A notation with ω is used to represent the stochastic nature of the matrices T and W and the vector h. The stochastic outcomes are denoted by $T(\omega)$, $W(\omega)$ and $h(\omega)$. After the stochastic outcome of $T(\omega)$, $W(\omega)$ and $h(\omega)$ are known, $y(\omega)$ has to be determined so that the second stage linear stochastic program $T(\omega)x + W(\omega)y(\omega) \geq h(\omega)$ is satisfied. This linear stochastic program describes the constraints that depend on the exogenous influences. The expected value of the second stage costs are included in the objective as $E_\xi[q^T y(\omega)]$, where q^T is the cost vector related to the second stage decision variables. Here ξ represents the probability space of $(T(\omega), W(\omega), h(\omega), y(\omega))$, for which the expected value of the second stage costs have to be taken. Finally non-negativity constraints are present in the model. The complete model reads as follows:

$$\begin{aligned} \text{Minimize} \quad & c^T x + E_\xi[q^T y(\omega)] \\ \text{s.t.} \quad & Ax \geq b \\ & T(\omega)x + W(\omega)y(\omega) \geq h(\omega), \\ & x \geq 0, \quad y(\omega) \geq 0. \end{aligned} \tag{6.1}$$

To attain a linear program, the discrete or discretized probability space ξ of ($T(\omega)$, $W(\omega)$, $h(\omega)$, $y(\omega)$) is written out to produce a very large linear program, called the *extensive form* of the stochastic program. The probability space ξ is represented by K possible outcomes: (T_k, W_k, h_k, y_k), with probabilities p_k, for $k = 1, ..., K$. This leads to the following large scale linear program:

$$\begin{aligned} \text{Minimize} \quad & c^T x + \sum_{k=1}^{K} p_k q_k^T y_k \\ \text{s.t.} \quad & Ax \geq b \\ & T_k x + W_k y_k \geq h_k, \text{ for } k = 1, ..., K, \\ & x \geq 0, \quad y_k \geq 0, \text{ for } k = 1, ..., K. \end{aligned} \tag{6.2}$$

Relation with the timetabling model The intuitive relation with our timetabling problem is given here. The first stage decision variables x represent the departure and arrival times in the timetable. In fact, the linear equations in $Ax \geq b$ represent the timetabling constraints. The costs of this timetable $c^T x$ can for example include excessive supplements or passenger transfer times. The second stage linear program $T_k x + W_k y_k \geq h_k$ has many similarities with the first stage problem. These equations make sure that the timetable is executed under disturbed circumstances. In the timetabling model, T_k is fixed, and it is equal for all stochastic outcomes $k = 1, ..., K$. This T-matrix is very similar to the A-matrix from the first stage, although cyclic relations have to be changed such that they link the repetitive cycles of the timetable with each other. Just as T is similar to A, h_k is related to b. To be more precise, h_k is more or less equal to b plus the disturbances. When the timetable restrictions $Tx \geq h_k$ are not satisfied because of large disturbances, this implies that delays y_k will occur in the execution of the timetable. These delays have a certain relation W_k with the other processes in the timetable. However, these relations are similar to the earlier timetabling relations given by A and T, and are also not stochastic. So $W_k = W$ for $k = 1, ..., K$. The delays are penalized with the cost vector q^T. The delays are penalized by a cost function, summed over all stochastic outcomes: $\sum_{k=1}^{K} p_k q^T y_k$.

Number of Realizations The k different outcomes, or realizations, are independent of each other in most stochastic programs. However, in our model these different realizations influence each other, due to the cyclicity of the timetable. At least, earlier cycles influence later cycles. Strictly following the model given by equations (6.2) this implies that all consecutive realizations of the timetable have to be formulated as one large stochastic outcome, with $K = 1$. This leads to one large matrix T, one large matrix W, one large vector y_k, and one large vector h_k. These matrices and vectors include all cycles. Note that this only leads to reliable outcomes if this one stochastic outcome of the disturbances is still a good representation of the probability space of the disturbances. However, instead of speaking of stochastic outcomes, we will speak of *realizations*. Each realization $z = 1, .., Z$ represents one cycle of the timetable. In Section 6.2, these realizations are independent of each other. In all later sections, the realizations are consequent cycles that depend on each other.

An alternative way of modeling our large number of realizations Z is breaking the interdependencies. For example, instead of a large number of cycles, the problem can be modeled as a sequence of days, with, say, 18 hourly cycles. Using this principle in the recourse model, $Z/18$ independent, but smaller, stochastic outcomes are used, without increasing the total size of the problem. This idea comes closer to the real world execution of passenger timetables, but it tells less about the mathematical stability. This alternative is not worked out further in this thesis.

6.2 Stochastic Optimization

The delay minimizing model described in this chapter is first applied to the supplement allocation problem addressed in Chapter 4. From hereon, the model from Section 4.3 is referred to as the numerical model, the model in this chapter is referred to as the stochastic model.

6.2.1 A Theoretical Example

In Sections 4.2 and 4.3, the supplement allocation for only one train is considered, so in these sections there is no interference between trains. Similar as in Chapter 4, consider one train traversing a number of consecutive track sections as in Figure 4.1.

The train incurs a disturbance on each of the trips. A given total amount of running time supplement has to be divided between the trips to minimize the average expected arrival delay for all stations. In other words, a timetable has to be found in which the supplements have been allocated optimally.

Before explaining the model any further, a few important assumptions for the problem are repeated.

- The full amount of supplement on a certain trip can be used to compensate a disturbance on the same trip. This is similar to assuming that the supplements are concentrated at the end of the trips.
- The timetable is not necessarily rounded to minutes.
- There are no cyclic dependencies, which means that the problem can be modeled on a linear time-axis.
- Trains start their journey without delay.
- Disturbances are small, and consequently subsequent cycles are independent.

Departures, arrivals, running times and supplements The model uses the running time supplements σr_n, for the sequential trips $n = 1, ..., N$ of the train line, as its key decision variables. Because the timetable is the basis for the stochastic model, the departure and arrival times are also decision variables in this model. Due to this structure, minimal running times mr_n and planned running times r_n are present in this model. The running time supplements σr_n are the differences between these two. Similarly, the difference between the planned dwell time s_n and the minimal dwell time ms_n equals the dwell time supplement σs_n:

$$\begin{aligned} \sigma r_n &= r_n - mr_n, \quad \text{for } n = 1, ..., N, \\ \sigma s_n &= s_n - ms_n, \quad \text{for } n = 1, ..., N-1. \end{aligned}$$

To avoid excessive planned travel times for the passengers, we want to restrict the total amount of supplement over the train line to σ_T:

$$\sum_{n=1}^{N} \sigma r_n + \sum_{n=1}^{N-1} \sigma s_n \leq \sigma_T. \tag{6.3}$$

However, just like in Chapter 4, it is assumed in this section that the dwell times are not disturbed and that the planned dwell times equal the minimal dwell times: $\sigma s_n = 0$ for all $n = 1, ..., N-1$.

The planned arrival time a_n equals the planned departure time at the preceding station, d_{n-1}, plus the planned running time:

$$a_n = d_{n-1} + r_n = d_{n-1} + mr_n + \sigma r_n, \text{ for } n = 1, ..., N, \tag{6.4}$$

where trip n with planned running time r_n runs from station $n-1$ to station n.

Besides the departure from station 0, a departure depends on the preceding arrival time and the dwell time.

$$d_n = a_n + s_n, \text{ for } n \text{ in } 1, ..., N-1. \tag{6.5}$$

Evaluation Now the realizations have to be formulated as well. In practice, each realization of a train run is different. Moreover, we want the constructed timetable to be the best timetable with respect to all realizations. Therefore, it is not sufficient to base the optimization on one realization. A large number of realizations is performed, using different stochastic disturbances for each of them. For each trip, these disturbances are randomly picked from the same non-negative distribution, but the distributions may vary over the trips. The disturbance on trip n in realization z is represented by δ_n^z for $n = 1, ..., N$ and $z = 1, ..., Z$, where Z is the number of realizations. The realized departure and arrival times at station n in realization z are given by $\tilde{d}_n^z$ and $\tilde{a}_n^z$, respectively.

The disturbances are determined before the model is run. This implies that the timetable obtained by the model is optimal with respect to these given disturbances. The predetermined disturbances can have any arbitrary distribution or size, since they are independent from the model.

Now it is also possible to calculate the delay propagation over the train line. The realized arrival times will never be earlier than planned. Furthermore, a realized arrival is at least as large as the realized departure time plus the minimal running time plus the running time disturbance.

$$\tilde{a}_n^z = \max\{a_n, \tilde{d}_{n-1}^z + mr_n + \delta_n^z\} \text{ for } n = 1, ..., N \text{ and } z = 1, ..., Z. \tag{6.6}$$

The realized departure times are also never earlier than planned. Additionally, a minimal dwell time has to be respected after the preceding arrival:

$$\tilde{d}_n^z = \max\{d_n, \tilde{a}_n + ms_n\} \qquad \text{for } n = 0, ..., N-1 \text{ and } z = 1, ..., Z, \tag{6.7}$$

Equations (6.6) and (6.7) are not linear, but they can be linearized easily. We obtain:

$$\begin{aligned}
&\tilde{a}_n^z \geq a_n, && \text{for } n = 1, ..., N \text{ and } z = 1, ..., Z,\\
&\tilde{a}_n^z \geq \tilde{d}_{n-1}^z + mr_n + \delta_n^z, && \text{for } n = 1, ..., N \text{ and } z = 1, ..., Z.\\
&\tilde{d}_n^z \geq d_n, && \text{for } n = 0, ..., N-1 \text{ and } z = 1, ..., Z,\\
&\tilde{d}_n^z \geq \tilde{a}_n^z + ms_n, && \text{for } n = 1, ..., N-1 \text{ and } z = 1, ..., Z.
\end{aligned} \tag{6.8}$$

Equations (6.8) are not exactly the same as equations (6.6) and (6.7), because $\tilde{a}_n^z$ and $\tilde{d}_n^z$ are not restricted from above. However, due to the minimization of the delays in the objective, this does not influence the results of the model.

The objective The objective is to minimize the unweighed average arrival delay over all realizations. The arrival delay Δ_i^z is the difference between the realized arrival time and the planned arrival time at station n in realization z:

$$\Delta_n^z = \tilde{a}_n^z - a_n, \text{ for } n = 1, ..., N \text{ and } z = 1, ..., Z. \tag{6.9}$$

Now the objective function can be given:

$$\textit{Objective} = \text{Minimize } \frac{1}{Z \cdot N} \sum_{z=1}^{Z} \sum_{n=1}^{N} \Delta_n^z. \tag{6.10}$$

Note that only the time differences between the departure and arrival variables are important in the model, and non-negativity constraints are not necessary for these variables. Any solution of the model can be shifted in time without altering the objective value. However, it is convenient to require the planned departure and arrival times to be non-negative. Furthermore, it is important to restrict the supplements to be non-negative for each trip. In the case of an integer timetable, one may allow the supplement on a trip to be slightly negative, because of the roundings, but that is not necessary here.

$$\begin{aligned}
&d_n \geq 0, \text{ for } n = 0, ..., N-1,\\
&a_n \geq 0, \text{ for } n = 1, ..., N,\\
&\sigma_n \geq 0, \text{ for } n = 1, ..., N.
\end{aligned} \tag{6.11}$$

The realizations for the arrival and departure times will then be non-negative automatically, because they are not allowed to be less than the planned departure and arrival times. The "realized supplements" cannot be negative, because the minimal running times are always respected for the realizations.

In Chapter 4 it was found that slight deviations in the amount of supplement per trip do not have a large influence on the average arrival delay, as long as the general shape of Figure 4.6. This implies that small stochastic influences can have a large impact on the optimal supplement allocation, although the objective is close to optimum. Still, a large number of realizations is needed to attain a reliable optimum. Section 6.4 provides some sampling possibilities to decrease the stochastic influence of the disturbances.

Without the integrality constraints for the departure and arrival times, the model given by equations (6.4), (6.5), and (6.8)–(6.11) is a linear programming model. However, with a large number of realizations, there are many constraints and decision variables, and the running time of the model will increase accordingly. Therefore, a small Z is preferred. On the other hand, a large number of realizations decreases the influence of the stochasticity of the model.

A small case The described model leads to many variables and constraints when the number of trips or the number of realizations is large. Here the model is written out for a small example with only two trips and two realizations: $N = 2$ and $Z = 2$. We call the stations A, B and C. The first trip runs from A to B, the second trip from B to C.

In order to simplify the example, we fix the planned departure time from station A to zero and the planned arrival time at station C to 16. The minimal running times from A to B and from B to C are both 7 minutes. The minimal dwell time and the planned dwell time at station B are both 1 minute. There are no dwell time disturbances ($\delta s_B^z = 0$). So there is no dwell time supplement, and 1 minute of running time supplement. This one minute has to be divided between the two trips, such that the average arrival delay over stations B and C over both realizations is minimal. The running time disturbances in the first realization are 1 minute on the first trip and 1 minute on the second trip; the disturbances in the second realization are 0 minutes on the first trip and 0 minutes on the second trip. The dwell times are not disturbed. The linear program resulting from this problem is given below. The timetabling constraints are given in the left column, while the constraints for the first and second realization are given in columns two and three, respectively.

After the objective on the first line, the minimal process times and the disturbances are given. Then the departure from A and the arrival at C are fixed in the first column, while the second and third column state that no realized departure or arrival can be earlier than planned. After that, the subsequent event times are related to each other for both the plan and the realizations. First, the arrival at B is linked to the departure from A, then the departure from B to the arrival at B, and finally the arrival at C to the departure from B. The arrival delays are determined with the last four equations.

$$\begin{array}{lll}
\text{Objective : minimize } \ \Delta, & & \Delta = \frac{1}{4}(\Delta_B^1 + \Delta_B^2 + \Delta_C^1 + \Delta_C^2) \\
 & & \\
mr_{A,B} = 7 & \delta_{A,B}^1 = 1 & \delta_{A,B}^2 = 0 \\
ms_B = 1 & \delta s_B^1 = 0 & \delta s_B^2 = 0 \\
mr_{B,C} = 7 & \delta_{B,C}^1 = 1 & \delta_{B,C}^2 = 0 \\
d_A = 0 & \tilde{d}_A^1 \geq d_A & \tilde{d}_A^2 \geq d_A \\
 & \tilde{a}_B^1 \geq a_B & \tilde{a}_B^2 \geq a_B \\
 & \tilde{d}_B^1 \geq d_B & \tilde{d}_B^2 \geq d_B \\
a_C = 16 & \tilde{a}_C^1 \geq a_C & \tilde{a}_C^2 \geq a_C \\
a_B = d_A + mr_{A,B} + \sigma_{A,B} & \tilde{a}_B^1 \geq \tilde{d}_A^1 + mr_{A,B} + \delta_{A,B}^1 & \tilde{a}_B^2 \geq \tilde{d}_A^2 + mr_{A,B} + \delta_{A,B}^2 \\
d_B = a_B + ms_B & \tilde{d}_B^1 \geq \tilde{a}_B^1 + ms_B & \tilde{d}_B^2 \geq \tilde{a}_B^2 + ms_B \\
a_C = d_B + mr_{B,C} + \sigma_{B,C} & \tilde{a}_C^1 \geq \tilde{d}_B^1 + mr_{B,C} + \delta_{B,C}^1 & \tilde{a}_C^2 \geq \tilde{d}_B^2 + mr_{B,C} + \delta_{B,C}^2 \\
 & \Delta_B^1 = \tilde{a}_B^1 - a_B & \Delta_B^2 = \tilde{a}_B^2 - a_B \\
 & \Delta_C^1 = \tilde{a}_C^1 - a_C & \Delta_C^2 = \tilde{a}_C^2 - a_C
\end{array} \tag{6.12}$$

It is easy to verify that, when this small problem is optimized, all supplements are allocated to the first trip: $\sigma_{A,B}^* = 1$ and $\sigma_{B,C}^* = 0$. The objective value, or average arrival delay, equals 0.25 minutes. A proportional allocation ($\sigma_{A,B} = \frac{1}{2}$ and $\sigma_{B,C} = \frac{1}{2}$) would lead to an average arrival delay of 0.38 minutes, 50% more than for the optimum.

Case: Total supplements equal to the total average disturbance In Section 4.3 a case is addressed, where the total amount of supplements equals the expected total disturbances. The same case is worked out here using the stochastic model.

Each trip of the train line incurs disturbances from the same exponential distribution with an average of one minute, i.e. $\lambda_n = 1$ for $n = 1, ..., N$. With N trips, this implies that N minutes of supplement are available. The objective is to minimize the average arrival delay by optimizing the supplement allocation over the trips.

Note that this case does not include any dwell time supplements: the available total supplement can be divided among the running times. As already noted in Section 4.3, the absolute value of the minimal and planned running times and minimal and planned dwell times are irrelevant to the model. The minimal running and dwell times may be chosen arbitrarily. Then the planned dwell times are equal to the minimal dwell times; the planned running times follow from the optimal supplement allocation found by the model.

Note that the model does not necessarily provide an integer timetable. To obtain an integer timetable one can round the arrival and departure times from the model to the nearest integer, possibly losing optimality. However, it is also possible to formulate the presented model as a mixed integer programming model, where $a_n \in \mathbb{N}$ and $d_n \in \mathbb{N}$ for $n = 1, ..., N$.

Results The case described above is implemented in the modeling system OPL Studio and solved with CPLEX on an Intel Pentium IV PC with a 3.0 GHz processor and 512 MB internal memory.

The model was used to optimize the described case for 2, 3, 4, 5, 6, 8, 10, 12, 15, 20 and 25 trips. The cases with 2 to 10 trips are worked out for 5,000 realizations, the cases with 12 to 25 trips for 1,000 realizations. Systematic sampling was applied to all these cases (see Section 6.4). For 10 trips there are 250,000 variables and 400,000 constraints, for 25 trips there are 125,000 variables and 200,000 constraints.

The results of the stochastic model are in line with the results of the numerical optimization in Section 4.3. Although there are small deviations from the earlier results, the same conclusions can be drawn from both models. Therefore, the conclusions are not repeated here. In Figures 6.1 and 6.2 the dark diamonds represent the results from the numerical optimizations in Section 4.3. The white diamonds represent the results of the stochastic model with 5,000 realizations for ten trips and 1,000 realizations for 25 trips.

Consequently, the resulting objective values are also almost the same for both models. For example, for 10 trips, the average delay is 1.049 in the numerical model and 1.046 minutes in the stochastic model. For 25 trips, the average arrival delays are 1.627 and 1.618 minutes, respectively.

The 10-trip situation with 1,000 realizations has been optimized for ten different disturbance samples from the same distribution. For these ten different samples, the estimated standard deviation of the objective, the overall average delay, is less than 0.01. The optimal supplement for the last trip equals 0 in all cases, which implies a standard deviation of 0. The standard deviation of the optimal supplement for the other trips ranges from 0.02 for the 1^{st} and 9^{th} trip, to 0.055 for the 5^{th} trip.

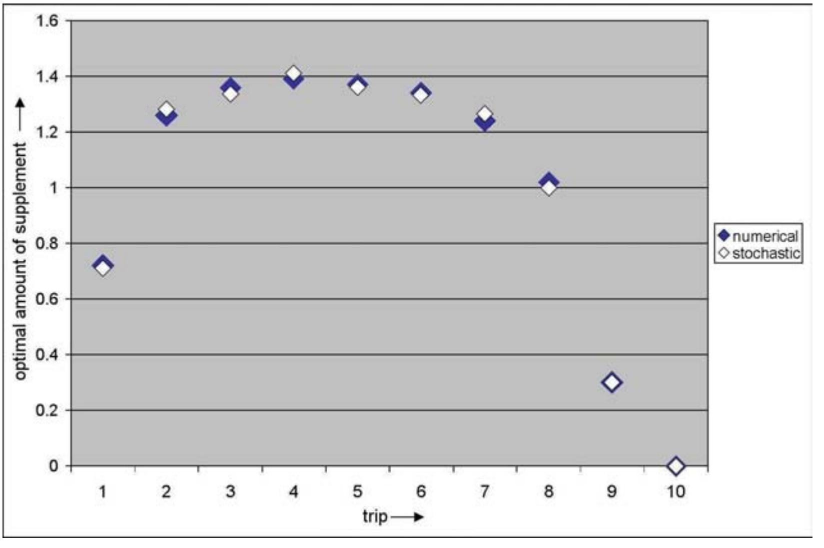


Figure 6.1: The optimal amount of running time supplement for 10 trips in both models.

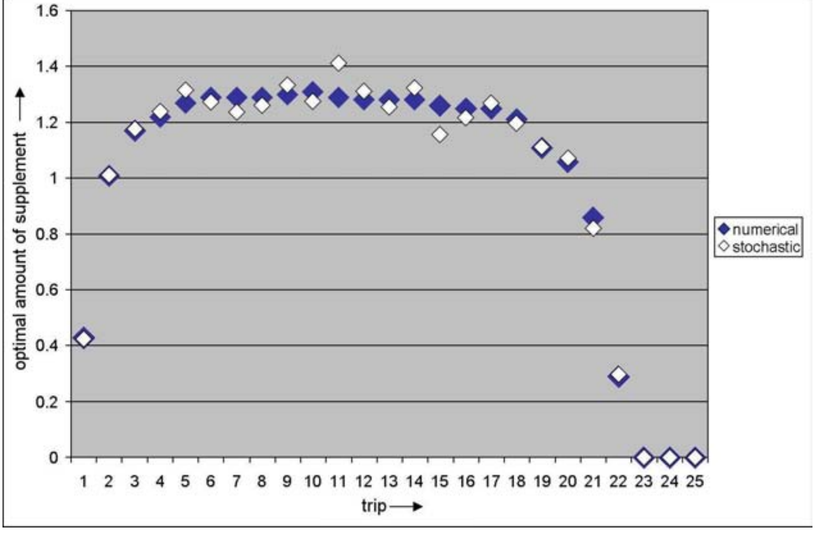


Figure 6.2: The optimal amount of running time supplement for 25 trips in both models.

Norms The current norm in the Netherlands for running time supplements is a proportional allocation of at least 7% of the minimal running time between each pair of *block points*, where a block point is a large station or an important junction. At those points, reliability is, intuitively, most important. The examples above show that a different allocation is more beneficial for small average disturbances. In the optimal situation, the first trips have relatively small supplements, and the last trips have no supplements at all. The middle trips have relatively large supplements. On average, the supplements shift to the front of the train line.

These results depend heavily on the disturbance-supplement ratio. When the disturbances are relatively large, the supplements should be shifted even further to the front. In that case, we also see a larger delay decrease. The opposite is true for relatively small disturbances. Furthermore, note that possible dwell time supplements increase the amount of the total supplements, which decreases the relative size of the disturbances. As a counterweight, the secondary delays not taken into account here, can be regarded as additional disturbances.

6.2.2 A Case: Haarlem–Maastricht/Heerlen

In this section we support the theoretical results from Section 6.2.1 with a practical case. This practical situation is identical to the situation described in Section 4.4: the Haarlem–Maastricht/Heerlen corridor. The precise description can be found there and is not repeated in this section. In this section, the same strategy is used as in Section 4.4: only the supplements of the 800- and 900-lines are optimized. In Section 6.5, the problem is modeled such that the supplements of the other lines and the buffer times between the trains can also be changed.

In this section, the supplements are optimized independently for the two lines and independently for the southbound and northbound directions. This is to compare the results of this section with the results in Section 4.4. This also implies that the planned running times are fixed. When the model prescribes a larger supplement for a certain trip, the minimal running time is implicitly decreased.

For the individual supplement optimization of the four lines (800-south, 800-north, 900-south, and 900-north) in this section, the stations Abcoude, Breukelen, Geldermalsen and Weert can be left out. This is because we do not consider interaction with other lines, and delays are not measured at these stations. In Section 6.5.2 the entire corridor is optimized. There we do take Abcoude, Breukelen, Geldermalsen and Weert into account.

The optimization of the individual lines can be done exactly as in the preceding section. The only difference is that there are different disturbance densities for the different trips. The model was optimized for 5,000 cycles and solved in about 30 minutes. The average disturbances per trip are given in the third column of Table 6.1 and are equal to the supplements in the reference situation in Section 4.4. This can be seen in the next column of the table.

trip	planned running time (min)	average disturbance (min)	running time supplements (min)		
			proportional 800 & 900	optimal 800	optimal 900
Hlm-Asd	14	1.03	1.03	0.85	0.87
Asd-Dvd	11	0.81	0.81	1.01	1.02
Dvd-Ut	17	1.25	1.25	1.43	1.44
Ut-Ht	28	2.05	2.05	2.63	2.67
Ht-Ehv	18	1.32	1.32	1.71	1.72
Ehv-Rm	31	2.27	2.27	2.57	2.64
Rm-Std	15	1.10	1.10	0.72	0.78
Std-Mt	15	1.10	1.10	0.00	-
Std-Hrl	18	1.32	1.32	-	0.00
Hrl-Std	18	1.32	1.32	-	0.98
Mt-Std	16	1.17	1.17	0.85	-
Std-Rm	15	1.10	1.10	1.24	1.28
Rm-Ehv	30	2.20	2.20	2.67	2.70
Ehv-Ht	19	1.39	1.39	1.85	1.85
Ht-Ut	28	2.05	2.05	2.59	2.58
Ut-Dvd	18	1.32	1.32	1.55	1.54
Dvd-Asd	12	0.88	0.88	0.39	0.37
Asd-Hlm	14	1.03	1.03	0.00	0.00

Table 6.1: The running time supplements for the proportional and the optimized situation.

Results The problem was again modeled in OPL Studio and solved with CPLEX. Table 6.1 shows the optimal supplement allocation according to the stochastic optimization.

After the numerical optimization of the supplements in Chapter 4, simulation was applied to attain the resulting delays. Although the secondary delays can also be calculated with the stochastic model, this is not worked out here. First, because the supplement allocation resulting from the stochastic model is so much similar to the results of the numerical model, that separate evaluation of the delay propagation is not necessary. Secondly, the stochastic model described here still needs several extensions for the interactions between trains. The interactions are explained easier for an entire corridor without known departure and arrival times. This is worked out for the whole Haarlem–Maastricht/Heerlen corridor in Section 6.5.

6.3 Convergence of the Stochastic Model

In this section we consider the same model as in Section 6.2.1, more specifically the case with two consecutive trips. The combined distribution of the disturbances δ_1 and δ_2 on these two trips has a finite set I of possible values. Each of these possible disturbance outcomes (δ_1^i, δ_2^i) has a probability of occurrence p^i, for all $i \in I$.

In this section we prove that the solution of the stochastic programming model converges to the optimal solution if the number of realizations goes to infinity. To prove this we assume that the optimal running time supplement allocation is unique. However, this is not essential. Further results related to the convergence of stochastic optimization solutions can be found in Linderoth et al. (2002).

6.3.1 Optimal running time supplement

As before, the total amount of running time supplement to be allocated equals σ_T. The running time supplement allocated to trip 1 is denoted by σ_1. Then the running time supplement allocated to trip 2 equals $\sigma_T - \sigma_1$. Figure 6.3 shows the partitioning of the positive (δ_1, δ_2) quadrant for a given value of σ_1 into four areas. $A_1(\sigma_1)$ is the area with relatively small disturbances on both trips. Here the supplements on both trips are large enough to compensate for the disturbances. $A_2(\sigma_1)$ is the area with relatively small disturbances on the first trip and relatively large disturbances on the second trip. This results in delays on the second trip only. $A_3(\sigma_1)$ is the area with relatively large disturbances on the first trip, and small disturbances on the second trip. This means that the first disturbance cannot be fully recovered by the running time supplement on the first trip. However, the supplement on the second trip is large enough to recover the remaining delay of the first trip plus the second disturbance. $A_4(\sigma_1)$ represents the area with relatively large disturbances on both trips. This implies arrival delays at the end of both trips.

With D_j^i we denote the delay of the train at the end of trip j ($j = 1, 2$) if the disturbances equal (δ_1^i, δ_2^i). In that case, the total weighted delay of the train over the two trips is denoted by Δ^i. For a given value σ_1 of the running time supplement on the first trip, the following (weighted) delays Δ^i are found:

- If (δ_1^i, δ_2^i) in $A_1(\sigma_1)$, then $\Delta_1^i = 0$ and $\Delta_2^i = 0$. Hence $\Delta^i = 0$.

- If (δ_1^i, δ_2^i) in $A_2(\sigma_1)$, then $\Delta_1^i = 0$ and $\Delta_2^i = \delta_2^i - (\sigma_T - \sigma_1)$. Hence $\Delta^i = w_2(\delta_2^i - \sigma_T + \sigma_1)$.

- If (δ_1^i, δ_2^i) in $A_3(\sigma_1)$, then $\Delta_1^i = \delta_1^i - \sigma_1$ and $\Delta_2^i = 0$. Hence $\Delta^i = w_1(\delta_1^i - \sigma_1)$.

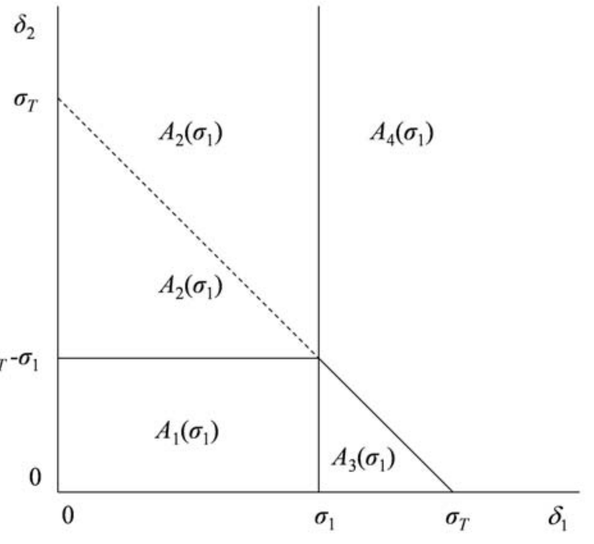


Figure 6.3: Partitioning into the areas $A_1(\sigma_1)$, $A_2(\sigma_1)$, $A_3(\sigma_1)$, and $A_4(\sigma_1)$.

- If (δ_1^i, δ_2^i) in $A_4(\sigma_1)$, then $\Delta_1^i = \delta_1^i - \sigma_1$ and $\Delta_2^i = \delta_1^i + \delta_2^i - \sigma_T$. Hence $\Delta^i = w_1(\delta_1^i - \sigma_1) + w_2(\delta_1^i + \delta_2^i - \sigma_T) = (w_1 + w_2)\delta_1^i + w_2\delta_2^i - w_1\sigma_1 - w_2\sigma_T$.

For a given value σ_1 of the running time supplement on the first trip, we can multiply these average delays per area with the probabilities p^i to obtain the average weighted delay $\Delta(\sigma_1)$ of the train:

$$\begin{aligned}\Delta(\sigma_1) = \quad & \sum_{i\in A_2(\sigma_1)} p^i w_2(\delta_2^i - \sigma_T + \sigma_1) + \sum_{i\in A_3(\sigma_1)} p^i w_1(\delta_1^i - \sigma_1) + \\ & \sum_{i\in A_4(\sigma_1)} p^i((w_1 + w_2)\delta_1^i + w_2\delta_2^i - w_1\sigma_1 - w_2\sigma_T). \qquad (6.13)\end{aligned}$$

Now the value σ_1^* for the running time supplement on the first trip has to be found, for which the average delay $\Delta(\sigma_1^*)$ is minimal.

The modification $\mu(\sigma_1)$ of the average delay on the two trips can be expressed in terms of a small modification μ_1 of the running time supplement on the first trip:

$$\begin{aligned}\mu(\sigma_1) \quad &= \sum_{i\in A_2(\sigma_1)} p^i w_2\mu_1 - \sum_{i\in A_3(\sigma_1)} p^i w_1\mu_1 - \sum_{i\in A_4(\sigma_1)} p^i w_1\mu_1 \\ &= \mu_1 \left(\sum_{i\in A_2(\sigma_1)} p^i w_2 - \sum_{i\in A_3(\sigma_1)\cup A_4(\sigma_1)} p^i w_1 \right)\end{aligned}$$

It follows that the average delay is minimal if the running time supplement σ_1 on the first trip is such that, around σ_1, the above expression changes from a negative value (decreasing average delay $\Delta(\sigma_1)$) to a positive value (increasing average delay $\Delta(\sigma_1)$). This implies that the expression

$$\sum_{i \in A_2(\sigma_1)} p^i w_2 - \sum_{i \in A_3(\sigma_1) \cup A_4(\sigma_1)} p^i w_1$$

is negative for values of σ_1 in the interval (σ_1^*-μ_1,σ_1^*), and positive for values of σ_1 in the interval (σ_1^*,σ_1^*+μ_1) for a sufficiently small value of μ_1.

Note that the value of the above expression only changes when one of the possible (δ_1, δ_2) combinations moves from one of the areas A_j to another. This means that the optimal value of the running time supplement σ_1^* on the first trip equals either a disturbance on the first trip $\{\delta_1^i \mid i \in I\}$, or the total amount of supplement minus a possible disturbance on the second trip $\{\sigma_T - \delta_2^i \mid i \in I\}$. Also note that here the assumption is used that there is a unique optimal allocation of the running time supplement. The foregoing implies that the average delay $\Delta(\sigma_1)$ is a convex piecewise linear function in σ_1.

6.3.2 Stochastic Optimization Model

Next, the described problem is solved with the stochastic optimization model for a random sample of Z realizations of pairs of disturbances. Let Z^i be the number of occurrences of the pair (δ_1^i, δ_2^i) in this sample.

Now, let σ_1 denote the proposed value for the running time supplement on the first trip. Then, in the same way as in the previous section, it follows that the average weighted delay $\Delta^Z(\sigma_1)$ can be expressed as follows:

$$\begin{aligned}\Delta^Z(\sigma_1) = \quad & \sum_{i \in A_2(\sigma_1)} \frac{Z^i}{Z} w_2(\delta_2^i - \sigma_T + \sigma_1) + \sum_{i \in A_3(\sigma_1)} \frac{Z^i}{Z} w_1(\delta_1^i - \sigma_1) + \\ & \sum_{i \in A_4(\sigma_1)} \frac{Z^i}{Z} ((w_1 + w_2)\delta_1^i + w_2\delta_2^i - w_1\sigma_1 - w_2\sigma_T). \end{aligned} \tag{6.14}$$

With an argument similar as in the previous section it can be shown that the optimal running time supplement σ_1^{Z*} on the first trip obtained by the stochastic optimization model is such that the expression

$$\sum_{i \in A_2(\sigma_1)} \frac{Z^i}{Z} w_2 - \sum_{i \in A_3(\sigma_1) \cup A_4(\sigma_1)} \frac{Z^i}{Z} w_1$$

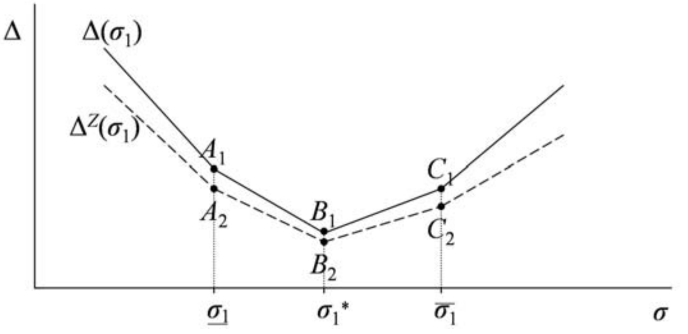


Figure 6.4: The convex piecewise linear functions $\Delta(\sigma_1)$ and $\Delta^Z(\sigma_1)$.

is negative for $\sigma_1 = \sigma_1^{Z*} - \mu_1$ and is positive for $\sigma_1 = \sigma_1^{Z*} + \mu_1$ for a sufficiently small value of μ_1.

The foregoing implies that the average delay $\Delta^Z(\sigma_1)$ is a convex piecewise linear function in σ_1. Parts of the graphs of the functions $\Delta(\sigma_1)$ and $\Delta^Z(\sigma_1)$ are represented in Figure 6.4.

6.3.3 Proof of Convergence

Now we claim that, with a probability tending to 1, the optimal value of the running time supplement on the first trip in the stochastic model (Section 6.3.2) equals the theoretical optimal value (Section 6.3.1) when the number of realizations goes to infinity.

Theorem 6.1. If the minimization problem has a unique optimal solution σ_1^* with $0 < \sigma_1^* < \sigma_T$, then $\lim_{Z\to\infty} P(\sigma_1^{Z*} = \sigma_1^*) = 1$.

Proof. Let $0 \leq \underline{\sigma_1} < \sigma_1^*$ be such that the interval $(\underline{\sigma_1}, \sigma_1^*)$ does not contain any value δ_1^i and such that the interval $(\sigma_T - \sigma_1^*, \sigma_T - \underline{\sigma_1})$ does not contain any value δ_2^i. Similarly, let $\sigma_1^* < \overline{\sigma_1} \leq \sigma_T$ be such that the interval $(\sigma_1^*, \overline{\sigma_1})$ does not contain any value δ_1^i and such that the interval $(\sigma_T - \overline{\sigma_1}, \sigma_T - \sigma_1^*)$ does not contain any value δ_2^i. Next, let μ_1 and μ_2 be defined by

$$\mu_1 := \sum_{i \in A_2(\underline{\sigma_1})} p^i w_2 - \sum_{i \in A_3(\underline{\sigma_1}) \cup A_4(\underline{\sigma_1})} p^i w_1$$

$$\mu_2 := \sum_{i \in A_2(\overline{\sigma_1})} p^i w_2 - \sum_{i \in A_3(\overline{\sigma_1}) \cup A_4(\overline{\sigma_1})} p^i w_1$$

Since σ_1^* is a unique minimum of the average delay $\Delta(\sigma_1)$, $\Delta_1 < 0$ and $\Delta_2 > 0$. Note that μ_1 and μ_2 are represented in Figure 6.4 by the differences $\Delta(\sigma_1^*) - \Delta(\underline{\sigma_1})$ and $\Delta(\overline{\sigma_1}) - \Delta(\sigma_1^*)$. In other words, the slopes of the solid lines A_1B_1 and B_1C_1 are negative and positive, respectively.

Next we will show that, if Z tends to infinity, then the probability that the differences $\Delta^Z(\sigma_1^*) - \Delta^Z(\underline{\sigma_1})$ and $\Delta^Z(\overline{\sigma_1}) - \Delta^Z(\sigma_1^*)$ are also negative and positive tends to 1. In other words, if Z tends to infinity, then the probability that the slopes of the dashed lines A_2B_2 and B_2C_2 in Figure 6.4 are negative and positive, respectively, tends to 1. A consequence is that, if Z tends to infinity, then the probability that $\sigma_1^{Z*} = \sigma_1^*$ tends to 1, as is to be proved.

To that end, first choose $\varepsilon > 0$ and let W be defined by $W := \max\{\ w_1,\ w_2\ \}$. Because of the Law of the Large Numbers (Feller, 1977), we know that for all $i \in I$ there exists an integer N_i such that for all $Z > N_i$ the following holds: $P\left(\left|\frac{Z^i}{Z} - p^i\right| \geq \frac{-\mu_1}{W|I|}\right) < \frac{\varepsilon}{2|I|}$. It follows that for all $Z > \hat{Z}_1 := \max\{\ N_i \mid i \in I\ \}$:

$$P\,(\text{the slope of } A_2B_2 < 0) =$$

$$P\left(\sum_{i\in A_2(\underline{\sigma_1})} \frac{Z^i}{Z} w_2 - \sum_{i\in A_3(\underline{\sigma_1})\cup A_4(\underline{\sigma_1})} \frac{Z^i}{Z} w_1 < 0\right) =$$

$$P\left(\sum_{i\in A_2(\underline{\sigma_1})} \left(\frac{Z^i}{Z} - p^i\right) w_2 - \sum_{i\in A_3(\underline{\sigma_1})\cup A_4(\underline{\sigma_1})} \left(\frac{Z^i}{Z} - p^i\right) w_1 < -\mu_1\right) \geq$$

$$P\left(\sum_{i\in I} \left|\frac{Z^i}{Z} - p^i\right| W < -\mu_1\right) = 1 - P\left(\sum_{i\in I} \left|\frac{Z^i}{Z} - p^i\right| W \geq -\mu_1\right) \geq$$

$$1 - \sum_{i\in I} P\left(\left|\frac{Z^i}{Z} - p^i\right| \geq -\frac{\mu_1}{W|I|}\right) > 1 - \frac{\varepsilon}{2}$$

Similarly, there exists an integer $\hat{Z}_2$ such that for all $Z > \hat{Z}_2$

$$P\,(\text{the slope of } B_2C_2 > 0) =$$

$$P\left(\sum_{i\in A_2(\overline{\sigma_1})} \frac{Z^i}{Z} w_2 - \sum_{i\in A_3(\overline{\sigma_1})\cup A_4(\overline{\sigma_1})} \frac{Z^i}{Z} w_1 > 0\right) > 1 - \frac{\varepsilon}{2}$$

As a consequence, for all $Z > \max\{\ \hat{Z}_1,\ \hat{Z}_2\ \}$ the minimum of $\Delta^Z(\sigma)$ is obtained for $\sigma_1^{Z*} = \sigma_1^*$ with probability at least $1 - 2(\frac{\varepsilon}{2}) = 1 - \varepsilon$. $\square$

We also claim that the difference between the objective value of the stochastic model and the real objective value tends to zero when the number of realizations tends to infinity.

Theorem 6.2. If the minimization problem has a unique optimal solution σ_1^* with $0 < \sigma_1^* < \sigma_T$, then for all $\delta > 0$ $\lim_{Z\to\infty} P(|\Delta(\sigma_1^{Z*}) - \Delta(\sigma_1^*)| < \delta) = 1$.

Proof. First, choose $\delta > 0$ and $\varepsilon > 0$, and let the positive number $\hat{Z}_0$ be such that $P(\sigma_1^{Z*} = \sigma_1^*) > 1 - \frac{\varepsilon}{2}$ for all $Z > \hat{Z}_0$. According to the proof of Theorem 1, such a number exists. Next, we have the following (in)equalities:

$$|\Delta^Z(\sigma_1^*) \; - \; \Delta(\sigma_1^*)| =$$
$$\left| \sum_{i\in A_2(\sigma_1^*)} \left(\frac{Z^i}{Z} - p^i\right) w_2(\delta_2^i - (\sigma_T - \sigma_1^*)) + \sum_{i\in A_3(\sigma_1^*)} \left(\frac{Z^i}{Z} - p^i\right) w_1(\delta_1^i - \sigma_1^*) + \right.$$
$$\left. \sum_{i\in A_4(\sigma_1^*)} \left(\frac{Z^i}{Z} - p^i\right) ((w_1 + w_2)\delta_1^i + w_2\delta_2^i \; - \; w_1\sigma_1^* - w_2\sigma_T) \right| \leq M \times \sum_{i\in I} \left|\frac{Z^i}{Z} - p^i\right|, \tag{6.15}$$

where M is an appropriately chosen positive number. Again, because of the Law of the Large Numbers, we know that for all $i \in I$ there exists an integer N_i such that for all integers $Z > N_i$ the following holds: $P\left(\left|\frac{Z^i}{Z} - p^i\right| \geq \frac{\delta}{M|I|}\right) < \frac{\varepsilon}{2|I|}$. Then it follows that for all integers $Z > \max\{\ N_i \mid i \in I\ \}$ the following holds:

$$\begin{gathered} P\left(\sum_{i\in I}\left|\frac{Z^i}{Z} - p^i\right| < \frac{\delta}{M}\right) = 1 - P\left(\sum_{i\in I}\left|\frac{Z^i}{Z} - p^i\right| \geq \frac{\delta}{M}\right) \geq \\ 1 - \sum_{i\in I} P\left(\left|\frac{Z^i}{Z} - p^i\right| \geq \frac{\delta}{M|I|}\right) > 1 - \frac{\varepsilon}{2}. \end{gathered} \tag{6.16}$$

Combining the results of (6.15) and (6.16), we find that the following holds for all integers $R > \max\{\ \{\ N_i \mid i \in I\ \} \cup \{\ \hat{Z}_0\ \}\ \}$:

$$P(|\Delta^Z(\sigma_1^{Z*}) - \Delta(\sigma_1^*)| < \delta) \geq P((|\Delta^Z(\sigma_1^{Z*}) - \Delta(\sigma_1^*)| < \delta) \cap (\sigma_1^{Z*} = \sigma_1^*)) =$$
$$P(|\Delta^Z(\sigma_1^{Z*}) - \Delta(\sigma_1^*)| < \delta) \mid \sigma_1^{Z*} = \sigma_1^*) \; \times \; P(\sigma_1^{Z*} = \sigma_1^*) \geq$$
$$P\left(\sum_{i\in I}\left|\frac{Z^i}{Z} - p^i\right| < \frac{\delta}{M}\right) \; \times \; P(\sigma_1^{Z*} = \sigma_1^*) \geq \left(1 - \frac{\varepsilon}{2}\right)\left(1 - \frac{\varepsilon}{2}\right) > 1 - \varepsilon.$$

□

Note that the results of Theorems 6.1 and 6.2 still hold if the optimal solution σ_1^* is not unique or if σ_1^* equals 0 or σ_T. However, slight modifications of the proofs are required then.

6.4 Reducing the Influences of the Disturbances

To reduce the stochastic influences of the disturbance sample, it may be helpful to sample in a smart way. Several techniques are known to reduce the variance of the model outcomes, by sampling in a structured manner. Some of these methods are described in this section. For further information, see for example Cochran (1977).

Antithetic Sampling To decrease the influence of the stochastic nature of the model, each realization can be repeated with antithetic disturbances. This means that if the stochastic disturbance δ_n^z for realization z is the x^{th} percentile of the cumulative disturbance distribution, then its counterpart $\delta_n^{z'}$ for the duplicate realization z' is the $100 - x^{\text{th}}$ percentile of the cumulative disturbance distribution function. This implies $F(\delta_n^z) + F(\delta_n^{z'}) = 1$. The number of realizations is $2 \cdot Z$.

Importance Sampling In importance sampling, the sample is not picked from the real, or estimated, density function $f(x)$, but from any arbitrary density function $g(x)$. The results of the sample element x are calculated and weighted by $f(x)/g(x)$ in the objective function of the problem. This method is particularly useful for densities $f(x)$ with low probability, but important tails. However, the weights for the sample elements are only correct when the different sample elements do not influence each other in the simulation. Therefore, this method is hardly applicable in our case.

Stratified Sampling Another way to decrease the stochastic influence of the sample is stratified sampling. This means that the sample population, in our case the disturbance distribution, is separated in several disjoint segments. Now we can pick randomly from each segment, taking into account that each segment is represented fairly in our sample. This is called *proportionate stratified* sampling. When a certain part of the population is relatively small, we can use *disproportionate stratified* sampling. This means that we take a relatively large sample from this small segment, but later weigh it to correct the results for the entire population.

Systematic Sampling Systematic sampling means that the sample contains every k^{th} element of the ordered population, where k = size of the population / sample size. In the case of a continuous distribution, every k^{th} percentile of the cumulative distribution function can be taken. Then k = 100/sample size, in our model $100/Z$. Additionally, the first element of the sample has to be chosen. Because the 0^{th} and 100^{th} percentile do not exist for some distributions, the $(\frac{100}{2Z})^{\text{th}}$ percentile is chosen as the first sample element. This implies that the last sample element, $Z-1$ elements later, is the $\frac{100}{2Z} + (Z-1) \cdot \frac{100}{Z} = (100 - \frac{100}{2Z})^{\text{th}}$ percentile. Both tails of the distribution that are left out of the sample have the same probability.

In short, the $\frac{z^*-1/2}{Z}$-th percentiles of the cumulative disturbance distribution of the appropriate trip n are taken, for $z^* = 1, ..., Z$ and $n = 1, ..., N$. Or equivalently: the cumulative distribution function $F_n(\delta_n^{z^*}) = \frac{z^*-1/2}{Z}$. Because the sample elements influence each other in the model, the disturbances $\delta_n^{z^*}$ have to be ordered randomly before they can be used. This ordering has to be done independently for all trips. The following procedure from Knuth (1997) can be used:

Input: δ_n^z = $\delta_n^{z^*}$, for $z = 1, ..., Z$ and $n = 1, ..., N$
$n = 1$
Do while $n \leq N$
    $R = Z$
    Do while $R \geq 2$
        Randomly pick an integer r from $1, ..., R$
        Swap δ_n^r and δ_n^R
        $R = R - 1$
    Loop
    $n = n + 1$
Loop
Output: δ_p^z, for $z = 1, ..., Z$ and $n = 1, ..., N$

Results Experiments with our optimization model show a 60% to 70% reduction in the standard deviation of the average delay with systematic sampling compared to simple random sampling. The standard deviation of the optimal running time supplement per trip decreased by approximately 20%. The standard deviation of the supplement is the highest for the fifth trip in the 10 trip case in Section 6.2.1. When optimizing over 1,000 cycles, we find a decrease of this standard deviation from 0.069 minutes to 0.055 minutes.

Systematic sampling is applied in Sections 6.2, 6.5, 6.6 and 6.8.

6.5 Optimization of a Corridor

6.5.1 The Model in General Terms

The real aim is to construct a timetable such that the overall average delay of the trains in the network is minimal. Besides the supplement allocation, buffer times between trains also play a key role in delay propagation. Therefore, the optimization of the supplements and buffers, and the calculations for the realizations of the whole network are to be integrated.

An important difference with the model so far is that the interdependencies between trains have to be modeled. Furthermore, the realizations are linked to each other: there is a 60-minute cyclic timetable.

The cyclic timetable For ease of presentation, we assume that the basic structure of the timetable is known a priori. That is, the train order of the trains on the considered tracks is known already, and the same holds for the layovers between the trains. This assumption implies that the main purpose of the model is to optimally allocate the buffer times to the headways and the supplements to the various running and dwell times. For further simplification of the model, we assume that the timetable does not contain any cyclic relations between trains, other than the periodicity of the timetable itself. Such cyclic relations could for example be due to the rolling stock circulation or to chains of passenger connections. With these assumptions, the model can be expressed on a linear time-axis, without the *modulo* Ts discussed in Section 3.1. Despite these assumptions, the constraint-graph will contain cycles, induced by the cyclicity of the timetable. All these assumptions can be relaxed, leading to a more complex model with many binary variables. How these assumptions can be relaxed is explained in Section 6.7.

We start with a general formulation of the model. The timetabling part is similar to the timetabling constraints in Section 3.1, but without the cyclicity notation. Furthermore, the running times are variable and the supplements have to be modeled.

The running time of the trip from station n to station n' of train t has minimal running time $mr^t_{n,n'}$, and can be modeled as

$$a^t_{n'} - d^t_n = mr^t_{n,n'} + \sigma r^t_{n,n'}, \tag{6.17}$$

where $\sigma r^t_{n,n'}$ is the running time supplement on this trip.

Likewise, the dwell time of train t at station n is modeled as

$$d^t_n - a^t_n = ms^t_n + \sigma s^t_n, \tag{6.18}$$

where ms_n^t is the minimal dwell time, and σs_n^t is the supplement. Passenger transfers and layover times can be modeled similarly.

Because the train order is assumed to be known, it is easy to determine which trains follow each other on a certain part of the infrastructure. Assume that train t' is the first train after train t, and that both trains use the same part of the infrastructure at departure from station n. The headway constraints for these trains t and t', can be formulated as:

$$h \leq d_n^{t'} - d_n^t. \tag{6.19}$$

Headways for arrival-arrival, arrival-departure, and departure-arrival combinations can be modeled in the same way.

Although the timetable is linearized, we still have to restrict all events at the same infrastructure to be planned within the cycle time T. Moreover, the timetable is supposed to be repeated, which means that a minimal headway has to be planned between the last train t' in one cycle, and the first train t in the next cycle on the same infrastructure. For the cyclic timetable this means that the last train t' is planned at most $T - h$ later than the first train t:

$$d_n^{t'} - d_n^t \leq T - h. \tag{6.20}$$

Further restrictions for the timetable construction are given in Section 3.1.

Other restrictions can also be included in the model. For example, one may want to restrict the total amount of supplements on a train to prohibit excessive travel times for the passengers. Let P^t be the sequence of running and dwell processes of train t. Then for all processes $p \in P^t$, σ_p^t is the supplement on that particular running or dwell time. Now the total supplement on the line can be restricted:

$$\sum_{p \in P^t} \sigma_p^t \leq \sigma_T^t, \text{ for all } t \in \Theta, \tag{6.21}$$

where σ_T^t is the maximal total amount of supplement for train t, and Θ is the set of trains.

Evaluation In the same way as in Section 6.2, the timetable is evaluated with regard to delay propagation during its construction. The main difference is that we do not have independent realizations, but Z consecutive periods that depend on each other. The realizations of the departure and arrival times of train t at station n are represented by $\tilde{d}_n^{t,z}$ and $\tilde{a}_n^{t,z}$, respectively, where z indicates the realization.

First of all, a train cannot depart nor arrive earlier than planned:

$$\begin{aligned} \tilde{d}_n^{t,z} &\geq d_n^t + z \cdot T, \quad \text{for } z = 1, ..., Z, \\ \tilde{a}_n^{t,z} &\geq a_n^t + z \cdot T, \quad \text{for } z = 1, ..., Z. \end{aligned} \tag{6.22}$$

All running times in all realizations can be disturbed. The exogenous disturbance of the running time of train t between stations n and n' in realization z is denoted by $\delta r_{n,n'}^{t,z}$. Dwell times can be disturbed likewise. A disturbance on the dwell time of train t at station n is denoted by $\delta s_n^{t,z}$. This leads to equations, linking the realized arrival and departure times with the minimal process times and the disturbances:

$$\begin{aligned} \tilde{a}_{n'}^{t,z} - \tilde{d}_n^{t,z} &\geq mr_{n,n'}^t + \delta r_{n,n'}^{t,z}, \quad \text{for } z = 1, ..., Z, \\ \tilde{d}_n^{t,z} - \tilde{a}_n^{t,z} &\geq ms_n^t + \delta s_n^{t,z}, \qquad \text{for } z = 1, ..., Z. \end{aligned} \tag{6.23}$$

Passenger transfers, layover times and headways can be disturbed similarly.

Although it can be easily incorporated in the model, we assume that the headways are not disturbed. Then we can model the headways for the realizations for two consecutive trains t and t' on the same infrastructure n as follows:

$$\tilde{d}_n^{t',z} - \tilde{d}_n^{t,z} \geq h. \tag{6.24}$$

Instead of the cyclicity headway restrictions in equation (6.20), the headway between the first train in cycle z has to be related to the last train in the preceding realization $z-1$ on the same infrastructure. Assume that train t is the last train using the particular part of the infrastructure n in cycle $z-1$, and t' is the first train in cycle z. If both trains have a departure at station n, this leads to the following constraint:

$$\tilde{d}_n^{t',z} - \tilde{d}_n^{t,z-1} \geq h, \text{ for } z = 2, ..., Z. \tag{6.25}$$

The arrival delay at station n of train t in realization z can be defined as

$$\Delta a_n^{t,z} = \tilde{a}_n^{t,z} - a_n^t - z \cdot T, \text{ for } z = 1, ..., Z. \tag{6.26}$$

Departure delays can be modeled identically. In any case, a set A of arrivals, and a set D of departures can be selected, for which we want to measure the delay:

$$\Delta = \frac{1}{(|A| + |D|) \cdot Z} \Big[\sum_{z=1}^{Z} \Big(\sum_{(t,n) \in A} \Delta a_n^{t,z} + \sum_{(t,n) \in D} \Delta d_n^{t,z} \Big) \Big], \tag{6.27}$$

where $|A|$ is the number of arrivals in the set A, and $|D|$ is the number of departures in the set D. The objective can now be defined as follows:

$$\mathit{Objective} = \text{Minimize } \Delta. \tag{6.28}$$

Besides delays, other timetable characteristics, such as running time supplements, can also be included in the objective function. Section 6.6 gives several examples of alternative objectives.

6.5.2 Extending the Case: Haarlem–Maastricht

The model described above was applied to the corridor from Haarlem to Maastricht and Heerlen, just like in Sections 4.4 and 6.2. Here we optimize the supplements and buffers in the timetable on this corridor, such that the average arrival delay at the ten measuring stations is minimal. The timetable is cyclic with a period of 60 minutes.

The trains are given in Figure 4.12, which also depicts the measuring points. The order of the trains on the tracks is already known and can be deduced from the cyclic 2004 timetable for the corridor which is described in Appendix B. This implies that the overtakings in Abcoude, Geldermalsen and 's-Hertogenbosch remain unchanged with respect to this reference timetable. Note that most lines run twice per hour, where each pair of trains on the same line are planned exactly 30 minutes apart.

The layover constraints at the line endpoints have not been modeled. This means that the southbound and northbound trains are almost independent from each other. Only level crossings exist at the south side of Sittard, where the tracks to Maastricht and Heerlen split. Because there is only a relatively small number of trains there, it is reasonable to keep this dependency outside the model. This implies that two independent problems are created: the southbound problem and the northbound problem. Only the southern direction is discussed here.

The planned running times in the tables of Appendix B include 7.92% of running time supplement for all trips. The only exceptions can be found in the 3500-line, which bears additional supplements of 1 minute between Duivendrecht (Dvd) and Abcoude (Ac), and vice versa, and 4 additional minutes between Abcoude and Breukelen (Bkl). Only the trains that are overtaken have dwell time supplements. The minimal planned dwell time for the 7300 in Abcoude and the 6000 in Geldermalsen (Gdm) is 4 minutes. This is because the overtaken train has to arrive at least 2 minutes earlier, and to depart at least 2 minutes later than the overtaking train. Because the 6000/9600 in 's-Hertogenbosch is overtaken by the 3500-line which dwells for one minute, there is a minimal planned dwell time of 5 minutes. All other minimal dwell times are equal to the planned dwell times.

When trains are not fully covered by the corridor, the departure and arrival times are, for the planning phase, fixed for the stations where those lines enter or leave the model. These arrival and departure times can be found in Table B.3.

The timetable of the reference situation in this section is also given in Appendix B. It is evaluated by the stochastic optimization model by fixing all departure and arrival times beforehand.

The disturbances are again assumed to be exponentially distributed. Each trip between two measuring stations is disturbed with an average equal to 7.92% of the minimal running time. This average is exactly equal to the planned supplements in the reference timetable.

Results for the Haarlem–Maastricht Corridor Again, the model is implemented in the modeling system OPL Studio and solved with CPLEX on an Intel Pentium IV PC with a 3.0 GHz processor and 512 MB internal memory. The optimization of the timetable for the Haarlem–Maastricht corridor and 500 realizations leads to a model with 160,000+ variables and 340,000+ constraints. It is solved in less than one hour.

The evaluation of the southbound reference timetable leads to an average arrival delay of 1.38 minutes. This is very close to the 1.39 minutes found by simulation in Section 4.4. The small deviation can be contributed to the stochasticity of both models and to the different ways of handling the trains. In the simulation, a first-come-first-serve method is applied, while in the stochastic optimization model train orders always remain unchanged during the realizations.

In the optimized timetable, the average arrival delay is reduced to 0.95 minutes, 31% less than the reference timetable. The 3-minute punctuality increased from 83.7% in the reference situation to 89.5% for the optimal timetable: this is a reduction of the number of late trains by 35%.

The running time supplement allocation which is found for the 800- and 900-line is given in the last column of Table 6.2. For the corridor optimization, an exact 30-minute service was enforced from Haarlem to Sittard, leading to identical supplements for the 800- and 900-line up to Sittard. Because of the longer running time from Sittard to Heerlen compared to Sittard-Maastricht, there is 0.22 minute more supplement available for the 900-line, which can only be assigned to Sittard-Heerlen. For the rest of the corridor, the supplement allocation looks much like the one found by the individual line optimization. The only remark that can be made in that respect is that slightly larger supplements are found for the most crowded part of the railway network between Amsterdam and Utrecht, and smaller supplements for the somewhat quieter network parts south of 's-Hertogenbosch.

trip	average disturbance (min)	optimal running time supplements (min)		
		optimized per line		optimized for the corridor
		800	900	800 & 900
Hlm-Asd	1.03	0.85	0.87	0.90
Asd-Dvd	0.81	1.01	1.02	1.16
Dvd-Ut	1.25	1.43	1.44	1.96
Ut-Ht	2.05	2.63	2.67	2.51
Ht-Ehv	1.32	1.71	1.72	1.55
Ehv-Rm	2.27	2.57	2.64	2.18
Rm-Std	1.10	0.72	0.78	0.67
Std-Mt	1.10	0.00	-	0.00
Std-Hrl	1.32	-	0.00	0.22

Table 6.2: The running time supplements for individual line optimization and entire corridor optimization.

6.5.3 Sensitivity Analysis for the Haarlem–Maastricht Corridor

The timetable found by the optimization model is only optimal with respect to the given disturbances. In this section we study the behavior of the optimized timetable for other sets of disturbances. For the analysis below, only the southbound timetable is evaluated. The *reference* timetable is the timetable with 7.92% running time supplements on all trips. The *preferred* timetable is the timetable which was optimal with respect to exponentially disturbed running times, with an average disturbance of 7.92% of the respective minimal running time.

First, we analyze the consequences of other sets of 500 realizations of random disturbances from the same disturbance distribution functions. The timetable is not optimized again for these different sets of disturbances. The departure and arrival times in the reference and preferred timetables are fixed, and the delay propagation is calculated for the new sets of disturbances. Ten random sets of disturbances from the same disturbance distributions were used, leading to ten evaluations of both models. This leads to the results given in Table 6.3. The spread of the average delay and the unpunctuality is at most 10%. This is relatively small in comparison to the differences between the reference and the preferred timetable. Note that the timetable is optimized for the average delays, and the punctualities are deduced from the results afterwards.

Another issue is that we do not know the real life disturbance distribution. Therefore, we again evaluate the preferred timetable, but now for sets of disturbances from

measure	reference timetable average delay	reference timetable unpunctuality	preferred timetable average delay	preferred timetable unpunctuality
average	1.34	15.6%	0.92	10.0%
minimum	1.30	15.0%	0.89	9.5%
maximum	1.38	16.3%	0.95	10.5%
range (in% of average)	6.0%	8.0%	5.8%	10.1%
standard deviation	0.028	0.45%	0.019	0.36%

Table 6.3: The influence of the randomness on the reliability measures.

other distributions. We describe the original disturbance distribution as an exponential distribution with average 1 instead of exponential with average 0.0792 times the minimal running time. By multiplying all distributions by 0.0792 times the minimal running time, we simplify the notations below.

First, the timetable is evaluated for exponential disturbances with other averages. Then it is evaluated for the situation where a large part of the running times is not disturbed, the rest is again exponentially disturbed. Furthermore, several uniform distributions are applied. Finally the timetable was submitted to triangular disturbances. The results are summarized in Table 6.4.

disturbance distribution	reference timetable average delay	reference timetable punctuality	preferred timetable average delay	preferred timetable punctuality	relative difference average delay	relative difference unpunctuality
exponential 0.50	0.19	99.1%	0.15	99.4%	21.4%	39.4%
exponential 0.75	0.64	93.8%	0.44	96.4%	30.3%	41.5%
exponential 1	**1.38**	**83.7%**	**0.95**	**89.5%**	**31.4%**	**35.2%**
exponential 1.25	2.42	71.7%	1.67	80.7%	31.0%	49.1%
exponential 1.50	3.68	60.5%	2.62	70.5%	28.7%	25.3%
50% 0 and 50% exp 1	0.54	94.3%	0.41	96.0%	25.4%	29.8%
50% 0 and 50% exp 1.5	1.43	83.6%	1.05	87.9%	26.4%	26.2%
50% 0 and 50% exp 2	2.77	70.7%	2.09	76.8%	24.6%	20.9%
80% 0 and 20% exp 2	0.83	90.3%	0.65	92.4%	22.4%	21.6%
80% 0 and 20% exp 3	2.04	79.2%	1.60	83.1%	21.5%	18.8%
80% 0 and 20% exp 4	3.71	67.7%	2.99	72.4%	19.5%	14.4%
uniform (0,2)	0.85	92.9%	0.51	97.1%	39.7%	59.0%
uniform (0,2.5)	1.66	79.8%	1.01	91.1%	39.1%	55.7%
uniform (0,3)	2.63	67.8%	1.68	80.9%	36.0%	40.7%
triangular (0,0,3)	0.91	91.9%	0.59	96.0%	34.8%	50.9%
triangular (0,0,4)	2.02	75.4%	1.33	85.3%	34.4%	40.2%
triangular (0,0,5)	3.43	59.4%	2.37	71.3%	30.9%	29.5%

Table 6.4: Reliability gain for different disturbance distributions. The parameter for the exponential distributions is the average (and not the reciprocal); the parameters for the triangular distribution are the minimum, the modus and the maximum.

This table shows that the preferred timetable is better than the reference timetable for a range of disturbance distributions. However, it can be observed that only few distributions lead to a larger gap in average delay and punctuality than the original exponential(1) distribution.

Finally, we studied the optimality of the preferred timetable with respect to other disturbance distributions. Therefore, the timetable was optimized again with the stochastic model, now with disturbances from the alternative distributions. This leads to a different optimal timetable for each row in Table 6.5. Again, note that the optimization is with respect to the average delay.

	preferred timetable		optimal timetable		relative difference	
disturbance distribution	average delay	punc-tuality	average delay	punc-tuality	average delay	unpunc-tuality
exponential 0.50	0.15	99.4%	0.11	99.6%	37.3%	50.8%
exponential 0.75	0.44	96.4%	0.42	96.8%	7.0%	11.9%
exponential 1	**0.95**	**89.5%**	**0.95**	**89.5%**	**0%**	**0%**
exponential 1.25	1.67	80.7%	1.64	80.9%	1.7%	2.7%
exponential 1.50	2.62	70.5%	2.53	71.2%	3.6%	2.5%
50% 0 and 50% exp 1	0.41	96.0%	0.39	96.3%	5.5%	8.7%
50% 0 and 50% exp 1.5	1.05	87.9%	1.04	88.0%	1.1%	1.1%
50% 0 and 50% exp 2	2.09	76.8%	2.03	77.2%	2.5%	1.3%
80% 0 and 20% exp 2	0.65	92.4%	0.63	92.6%	1.9%	3.3%
80% 0 and 20% exp 3	1.60	83.1%	1.54	83.7%	3.9%	3.4%
80% 0 and 20% exp 4	2.99	72.4%	2.81	73.6%	6.2%	3.9%
uniform (0,2)	0.51	97.1%	0.48	97.9%	7.3%	41.5%
uniform (0,2.5)	1.01	91.1%	1.00	91.2%	1.6%	1.5%
uniform (0,3)	1.68	80.9%	1.63	81.3%	3.4%	2.3%
triangular (0,0,3)	0.59	96.0%	0.58	96.4%	2.8%	9.6%
triangular (0,0,4)	1.33	85.3%	1.31	85.3%	1.2%	0.2%
triangular (0,0,5)	2.37	71.3%	2.26	72.6%	4.8%	7.1%

Table 6.5: Comparing the preferred and the optimal timetables for different disturbance distributions.

Table 6.5 shows that the preferred timetable is close to the optimum for a range of disturbance distributions.

6.6 Alternative Objectives

Besides minimizing the average delay, one can think of many other criteria which are worthwhile to be optimized. Two of them are discussed in this section. First, another reliability measure is optimized: the punctuality. Later, the passenger travel time is included in the objective function. For this new objective, a case is worked out considering passenger transfers in Amersfoort.

The objectives in this section are based on the assumptions and restrictions of the stochastic optimization model that is described in Section 6.5.1.

6.6.1 Maximizing Punctuality

In practice, punctuality is often used as an indication for reliability in public transport. As a reminder, punctuality is the percentage of trains arriving *on time*. In this case, all trains arriving at most the punctuality margin later than planned are considered to be on time.

To determine the punctuality, the delay of the trains has to be compared to the punctuality margin D. This implies that for all appropriate arrivals (all $a \in A$) in all cycles ($z = 1, ..., Z$) a binary variable π_a^z has to be introduced. The variable π_a^z equals 1 when arrival a in realization z is completed within margin D from its planned completion time (it is considered punctual), and 0 otherwise. On top of the other constraints, this can be enforced by the following restrictions:

$$\begin{aligned} &\pi_a^z \in \{0, 1\}, && \text{for } a \in A \text{ and } z = 1, ..., Z, \\ &M(1 - \pi_a^z) \geq \Delta_a^z - D, && \text{for } a \in A \text{ and } z = 1, ..., Z, \end{aligned}$$

where M is a sufficiently large parameter, and Δ_a^z is the delay of arrival a in realization z. The right hand side of the second equation is positive when the delay Δ_a^z is larger than the punctuality margin D. This implies that $1 - \pi_a^z$ is also larger than zero: $\pi_a^z = 0$.

Now the punctuality Π is obtained easily:

$$\Pi = \frac{1}{|A| \cdot Z} \sum_{a \in A} \sum_{z=1}^{Z} \pi_a^z.$$

And the objective is replaced by

$$\mathit{Objective} = \text{Maximize } \Pi.$$

Although this alternative objective is modeled relatively easily, $|A| \cdot Z$ binary variables are introduced in the model. This is computationally difficult to solve.

6.6.2 Minimizing Travel Time

Besides the reliability, the planned travel times for the passengers provide an important characteristic of the timetable. Earlier, travel times were restricted by equation (6.21), which bounds the supplement on a process or set of processes. However, the bounds on the supplements have to be set beforehand, despite the fact that the optimal amount of supplement is unknown beforehand. Therefore, we prefer to leave equation (6.21) out of the model, and insert the travel times into the objective function.

Note that realized travel times equal the planned travel times plus the realized delay. However, minimizing realized travel times only, does not make much sense: if feasible, no supplements would be planned. In that case, a train can never make up for the incurred delays, and the timetable would be quite unreliable. Therefore, we create a multi-criteria objective. This objective weighs the reliability, in terms of average delay, against the planned travel times.

Weighing factors Besides weighing delays against travel times, we also want to weigh the delays against each other, and the travel times against each other. Delays and travel times of trains with many passengers are more important than those of trains with only a few passengers. In this section first the weighted average delay is calculated, where the weights represent the expected numbers of passengers. Then the weighted planned travel time is determined. Here the weights, again, represent the expected numbers of passengers, but have to be determined differently, as will be explained later. Finally, the two are weighed against each other. The weights used here come from other research, in particular research on travel time and delay perception of passengers (Rietveld et al., 1998).

The weighted average delay is calculated, by weighing every arrival delay for the number of passengers that has this trip as its last trip: α_a, for all $a \in A$. Intermediate delays are not important for the passengers. This is even more true for our model than in reality, because all (defined) passenger transfers are upheld in the model. We assume that the number of passengers for a trip is equal throughout the realizations. However, it is easy to implement different passenger counts for different realizations to, for example, mimic the peak hours. Now the weighted average arrival delay can be defined as:

$$\Delta^\alpha = \sum_{a \in A} \sum_{z=1}^{Z} (\alpha_a \cdot \Delta_a^z) \Big/ Z \sum_{a \in A} \alpha_a, \tag{6.29}$$

where the divisor is the total number of passengers in Z realizations.

The other part of the multi-criteria objective is the planned travel time. The passenger travel time consists of a sequence of running times, dwell times and transfer times. The process times of this whole sequence have to be counted. Let B be the set of all processes that are part of the journey of at least one passenger. Then β_b is defined as the number of passengers using process b for their journey. Note that minimizing the travel time is equivalent to minimizing the supplements in the travel components. Then we define the average planned travel time supplements J:

$$J = \sum_{b \in B} (\beta_b \cdot \sigma_b) \Bigg/ \sum_{a \in A} \alpha_a, \tag{6.30}$$

where σ_b is the planned supplement on process b. The divisor is the total number of passengers in one realization.

However, one can argue that the passenger perception of the travel time is not equal for the different travel components. For example, waiting for a connection at a windy station, the transfer time, is worse than sitting in a dwelling train. Time spent in a running train is perceived even better. This can be modeled by dividing the processes in the set B into three disjoint process sets B_r, B_s, and B_{pt}, for the planned running times, dwell times and passenger transfer times, respectively. The weights w_r, w_s, and w_{pt} represent the perceived duration weights of these processes. The average perceived planned travel time J^w is then defined as:

$$J^w = \left[w_r \sum_{b \in B_r} (\beta_b \cdot \sigma_b) + w_s \sum_{b \in B_s} (\beta_b \cdot \sigma_b) + w_{pt} \sum_{b \in B_{pt}} (\beta_b \cdot \sigma_b) \right] \Bigg/ \sum_{a \in A} \alpha_a. \tag{6.31}$$

So far, we have found the average delay per passenger and the perceived travel time per passenger. Earlier research (Rietveld et al., 1998) showed that passengers consider a minute of delay more costly than a minute of planned travel time. Define w_Δ as the weight indicating the importance of the delays with respect to the travel time. Now the timetable can be optimized without dictating the total amount of supplements in the system. The model is the same as in Section 6.5, but equation (6.21) can be skipped and the objective (6.28) has to be replaced by:

$$\textit{Objective} = \text{Minimize } J^w + w_\Delta \cdot \Delta^\alpha.$$

The parameter w_Δ can be viewed as the risk aversion of passengers. When $w_\Delta \leq 1$ this implies *risk loving* passengers. This would lead to a timetable without any supplements. Rietveld et al. (1998) finds money values for both travel times and delay times. Dividing these two leads to $w_\Delta = 2.37$.

6.6.3 Optimizing Passenger Transfers: the Amersfoort Case

The Dutch railway system is characterized by many connecting services to provide good connections between a wide range of origins and destinations. Connections have to be planned long enough to enable passengers to alight from the feeder train, and to walk to and board the connecting train. To offer a reliable transfer, some supplement should be included in the transfer time. This can be used to absorb a small delay of the feeder train.

In the Netherlands, passenger transfers are often only two to six minutes in the timetable. Two minutes is really a lower bound, even when the trains dwell at opposite sides of the same platform. Notably in other countries connection times are often longer, ranging from about 5 to 20 minutes. The tightness of the connections in the Netherlands is a non-negligible source of delay propagation, foremost because connecting trains often wait for passengers of delayed feeder trains.

The case described here concerns one connection only. However, many similar connections can be modeled simultaneously when a larger network is modeled.

The size of the passenger transfer problem that is considered, has been kept small for ease of presentation. Here we are looking at the cross-platform connection between the 700-line and 1700-line in Amersfoort (Amf). A *cross-platform connection* is a connection, where both involved trains are along opposite sides of the same platform to facilitate the transfer in both directions. In reality, the 700-line is coupled in Amersfoort: one train from Schiphol and one train from Amsterdam continue their trip together towards Zwolle and further. For ease of presentation we assume that the 700-line is a regular line running from Amsterdam (Asd) via Amersfoort and Zwolle (Zl) to Groningen; the 1700-line runs from The Hague via Utrecht (Ut), Amersfoort and Apeldoorn (Apd) to Enschede. However, we only consider the Amsterdam–Amersfoort–Zwolle part of the 700-line, and the Utrecht-Amersfoort-Apeldoorn part of the 1700-line.

The trains from the 700- and 1700-line are scheduled to be along opposite sides of the same platform in Amersfoort at the same time. This facilitates journeys from Amsterdam to Apeldoorn and from Utrecht to Zwolle. In the stochastic model all passenger transfers are attained in the realizations. This implies that when one of the two trains is delayed by more than the transfer supplement, it will also delay the departure of the other train.

The optimization involves six arrival and departure times. For both the 700- and 1700-line the arrival and departure times in Amersfoort are determined. Furthermore the arrival times of the 700-line in Zwolle and the 1700-line in Apeldoorn are optimized. The arrival and departure times in Amersfoort determine the planned dwell times and planned transfer times there. This directly defines the available supplements in Amersfoort. The departures from Amersfoort and the arrivals in Zwolle and Apeldoorn determine the planned running times and the available running time supplements for the Amersfoort–Zwolle and Amersfoort–Apeldoorn trips. The minimal dwell time of both trains is assumed to be 1.5 minutes, the minimal transfer time from one train to the other is 2 minutes. The minimal running times are 33.6 minutes for Amersfoort–Zwolle and 22.9 minutes for Amersfoort–Apeldoorn.

The arrival delay distributions of the 700- and 1700-line in Amersfoort are derived from their respective real life 3-minute punctuality. In real life, the arrival punctuality in Amersfoort for the 700-line is 80.2%, for the 1700-line 83.8%. The assumption that the arrival delays are exponentially distributed leads to the equations $1 - e^{-3\lambda^{700}_{Amf}} = 0.802$ and $1 - e^{-3\lambda^{1700}_{Amf}} = 0.838$. Solving these gives $\lambda^{700}_{Amf} = 0.54$ and $\lambda^{1700}_{Amf} = 0.61$, or average arrival delays of 1.85 and 1.65 minutes, respectively. Furthermore, the disturbances on the running times on Amersfoort–Zwolle and Amersfoort–Apeldoorn have to be defined. Again exponential distributions are chosen, now with an average of 5% of the minimal running time. This is 1.68 minutes on average for Amersfoort–Zwolle and 1.14 minutes for Amersfoort–Apeldoorn.

One can recognize four types of passengers in this case. These different types are categorized to be able to weigh the different processes and delays in the objective function. The weights represent the number of passengers for the respective journey. The number of passengers traveling from station n to station n' is given by $\gamma_{n,n'}$.

- **Amersfoort as final destination** These passengers can be left out of the model, because they have a given arrival delay distribution.

- **Amersfoort as starting point** For these passengers, both the planned running time and the arrival delay are weighed. These passengers are represented by $\gamma_{Amf,Zl}$ and $\gamma_{Amf,Apd}$.

- **Direct journeys through Amersfoort** Direct passengers are those from Amsterdam to Zwolle ($\gamma_{Asd,Zl}$) and those from Utrecht to Apeldoorn ($\gamma_{Ut,Apd}$). For these passengers the planned dwell time in Amersfoort, the planned running time from Amersfoort to their destination, and the delay at the end of their journey are weighed.

- **Journeys with transfer in Amersfoort** Passengers from Amsterdam to Apeldoorn ($\gamma_{Asd,Apd}$) and passengers from Utrecht to Zwolle ($\gamma_{Ut,Zl}$) have to transfer in Amersfoort. For them, the planned transfer time, the planned running time from Amersfoort to the destination, and the arrival delay are weighted.

Instead of using the planned process times, the process time supplements are used in the objective. This does not influence the optimal departure or arrival times. When we weigh each of the process supplements and the average delays for the number of passengers, this leads to the following objective function:

$$
\begin{aligned}
\textit{Objective} = \text{Minimize} \quad & \gamma_{Amf,Zl} \cdot (\sigma r_{Amf,Zl} + w_{\Delta} \cdot \Delta_{Zl}^{700}) + \\
& \gamma_{Amf,Apd} \cdot (\sigma r_{Amf,Apd} + w_{\Delta} \cdot \Delta_{Apd}^{1700}) + \\
& \gamma_{Asd,Zl} \cdot (\sigma s_{Amf}^{700} + \sigma r_{Amf,Zl} + w_{\Delta} \cdot \Delta_{Zl}^{700}) + \\
& \gamma_{Ut,Apd} \cdot (\sigma s_{Amf}^{1700} + \sigma r_{Amf,Apd} + w_{\Delta} \cdot \Delta_{Apd}^{1700}) + \\
& \gamma_{Asd,Apd} \cdot (\sigma pt_{Amf}^{700,1700} + \sigma r_{Amf,Apd} + w_{\Delta} \cdot \Delta_{Apd}^{1700}) + \\
& \gamma_{Ut,Zl} \cdot (\sigma pt_{Amf}^{1700,700} + \sigma r_{Amf,Zl} + w_{\Delta} \cdot \Delta_{Zl}^{700}),
\end{aligned}
$$

where σr, σs, and σpt are the supplements on running time, dwell time and passenger transfer, respectively. This objective function can be viewed as the perceived travel time loss.

The relative number of passengers (over the whole day) for each of the six possible journeys in the model are obtained from the Marketing Research & Consultancy department of NS. These numbers are $\gamma_{Asd,Zl} = 224$, $\gamma_{Asd,Apd} = 91$, $\gamma_{Ut,Zl} = 229$, $\gamma_{Ut,Apd} = 310$, $\gamma_{Amf,Zl} = 69$, and $\gamma_{Amf,Apd} = 76$. Rietveld et al. (1998) found values for travel times and delays in several transport modes. The $w_{\Delta} = 2.37$ is the fraction of these two for public transport.

Results The model is solved for 10,000 realizations. The disturbance samples are systematic and the model is solved within a few minutes. The resulting arrival times (a) and departure times (d) in Amersfoort, Zwolle and Apeldoorn are given in Table 6.6.

700	2004	optimal	1700	2004	optimal
Amersfoort a	06.0	06.8	Amersfoort a	07.0	08.0
Amersfoort d	10.0	10.0	Amersfoort d	10.0	11.0
Zwolle a	46.0	47.3	Apeldoorn a	35.0	36.0

Table 6.6: The optimal timetable for the 700- and 1700-line. In minutes past each hour.

However, Table 6.7 provides more interesting figures. First of all, the objective value of the optimal timetable is 6.14 minutes, compared to 6.21 minutes for the 2004 timetable of NS. This implies that the optimized connection and running times decrease the perceived travel time loss by only 1.1%.

The perceived travel time loss decreased, despite an increase of planned running times. This decrease is caused by a decrease of planned dwell and transfer times. The average arrival delay is also smaller. In particular, the dwell time of the 700-line is reduced in Amersfoort. This is accompanied by a shorter transfer from the

1700- to the 700-line. However, the running time supplement on the 700-line between Amersfoort and Zwolle is increased. Especially the arrival delay in Zwolle has been reduced.

in minutes	2004	optimal
perceived travel time loss	6.21	6.14
dwell time supplement 700	2.5	1.7
dwell time supplement 1700	1.5	1.5
transfer time supplement 700-1700	2.0	2.3
transfer time supplement 1700-700	1.0	0.0
running time supplement Amf-Zl	2.4	3.7
running time supplement Amf-Apd	2.1	2.1
average arrival delay Zl	1.17	0.99
average arrival delay Apd	0.88	0.84

Table 6.7: Supplements and delays for the real life and optimal timetables around Amersfoort.

Note that it is intuitively correct to have more supplements on the running times instead of on the dwell times, because we do measure arrival delays and not departure delays.

Canceling Passenger Transfers During Operations In the model above it is assumed that trains always wait for each other to uphold passenger transfers. This is easily ensured by implementing passenger transfer restrictions for each realization, which are similar to those for the planned timetable.

Laisser-faire is the opposite policy, where the trains do not wait for each other to uphold a connection. The passenger connection is only upheld if the delay of the feeder train is less than the delay of the connecting train plus the transfer supplement. This policy is also implemented easily: the passenger transfers are present in the timetabling part of the model, but not in the realizations.

However, in reality trains will sometimes wait for each other, and sometimes the transfer will be canceled. In fact, there is a trade-off between missing the transfer for the transferring passengers on one hand, and the extended travel time of the passengers that are already in the train. Goverde (1998) describes this problem in detail. For exponentially distributed disturbances, he analytically optimizes the optimal planned transfer time and the optimal waiting time for a single passenger transfer.

In fact, in the Netherlands maximal waiting times are defined for passenger transfers. This is the maximal time that a connecting train will wait to provide a transfer possibility for passengers from another train. Although it is relatively easy to implement a conditional waiting strategy, there is an important drawback.

Consider the passenger transfer from train t to train t'. For ease of presentation we assume that there is no transfer time supplement. The operational strategy for train t' is to wait if the delay of the arrival of feeder train t is less than the waiting time wt. The connecting train t' has to wait to provide a connection in the case that the delay of the feeder train, Δ^f, is less than wt. This means that the connecting train incurs a secondary delay equal to that of the feeder train $\Delta^c = \Delta^f$. In fact this may be profitable for the overall solution, but not necessarily, as is explained in the following.

The trade-off here is between the delay for the transferring passengers when they miss the connection, and the delay that the other passengers incur because they have to wait for the transferring passengers. Now recall the following two characteristics of the model: (i) the disturbance on the next trip is not known in real operations, so the decision should only depend on the situation in the connection station and not on what happens later. However, the disturbances are defined a priori in the model; (ii) within the current model it is possible to cancel the connection by delaying the feeder train a little bit more. This is possible because the arrival time is based on a set of inequalities, which only define a lower bound for the arrival time. This may be beneficial, when the connecting service will incur large disturbances after leaving this station.

To avoid the unnecessary delays, it has to be made sure, that the arrival is exactly equal to the maximum completion time of the underlying processes. This can be realized by introducing a large number of binary variables and constraints including *big-M* parameters. The waiting time rule itself also requires binary variables and constraints with *big-M* parameters. The same conclusions can be drawn for other operational decision rules. Therefore it is recommended to avoid these operational rules in the model when possible.

6.7 Relaxing the Assumptions

In the preceding sections, the timetable was subject to several assumptions. These assumptions are necessary to model the timetable without binary variables. This is computationally an advantage, because our stochastic optimization model is a linear programming model in that way. However, if desired, we can relax these assumptions. This enables us for example to model a more coherent network with cyclic relations. Furthermore, we do not necessarily have to provide the model with the train orders on the tracks.

6.7.1 Cyclic Relations

Chains of passenger transfers and rolling stock layovers lead to cyclic relations in a periodic timetable. The presence of these cycles implies that not all relations can be formulated within one cycle. For some of these relations it is clear how many cycles there are between the two events leading to a cyclic relation. For other relations that is unknown a priori.

Fixed offset Consider a train t from station n to station n' and a train t' in the opposite direction. The rolling stock used for these trains is connected to each other at stations n and n'. The minimal layover times are $ml_{n'}^{t,t'}$ and $ml_n^{t',t}$. Remember that in the linear form, which was used up to this section, the arrival of a train is always later than its departure. Now the layover at station n can be formulated as

$$d_n^t \geq a_n^{t'} + ml_n^{t',t}. \tag{6.32}$$

However, the layover at the other end-point cannot be formulated likewise: the sequence $d_{n'}^{t'}, a_n^{t'}, d_n^t, a_{n'}^t$ is increasing. Due to the periodicity of the timetable, the planned departure $d_{n'}^{t'}$ can also be related to an 'earlier' arrival $a_{n'}^{t'}$. To do so, a time offset is necessary, which is a multiple of the period of the timetable.

Assume that it is known that there are $k_{n'}^{t,t'}$ cycles between two consecutive departures of the same train composition from station n'. Then the cyclic relation can be formulated easily:

$$d_{n'}^{t'} \geq a_{n'}^t + ml_n^{t,t'} - k_{n'}^{t,t'} \cdot T. \tag{6.33}$$

And for the realizations we can write:

$$\begin{aligned} \tilde{d}_n^{t,z} &\geq \tilde{a}_n^{t',z} + ml_n^{t',t}, && \text{for } z = 1, ..., Z, \\ \tilde{d}_{n'}^{t',z} &\geq \tilde{a}_{n'}^{t,z-k_n^{t,t'}} + ml_{n'}^{t,t'}, && \text{for } z = k_{n'}^{t,t'} + 1, ..., Z. \end{aligned} \tag{6.34}$$

It has to be noted that in a case with more lines and more relationships, it may be necessary to use time offsets for both layover times.

Variable offset Unfortunately, it is not always known beforehand how many cycles elapse between two similar events in different realizations. Then $k_{n'}^{t,t'}$ becomes an integer variable. However, most of the time a good approximation of $k_{n'}^{t,t'}$ can be given, due to the fact that a difference in $k_{n'}^{t,t'}$ of 1 implies a difference of a full period. In other words, with an hourly pattern, an increase of $k_{n'}^{t,t'}$ by 1 means that the time needed for all processes in the cyclic chain increases by 60 minutes. For ease of presentation, we therefore assume that we know that the time offset is either $k_{n'}^{t,t'}$

or $k_{n'}^{t,t'} + 1$ times the timetable period. To model this, we assume again that $k_{n'}^{t,t'}$ is a parameter and $K_{n'}^{t,t'}$ is a binary variable. For the plan, equation (6.33) can easily be rewritten as

$$d_{n'}^{t'} \geq a_{n'}^{t} + ml_{n'}^{t,t'} - (k_{n'}^{t,t'} + K_{n'}^{t,t'}) \cdot T. \tag{6.35}$$

This leads to the complication, that $K_{n'}^{t,t'} = 1$ is always easiest to satisfy for the model. Therefore, the K-variables have to be restricted in the model, because, in case of rolling stock layovers, every increment of K costs additional rolling stock. The cycles can also be caused by for example passenger transfers. In that case, every increment of K increases the travel time of transferring passengers.

The K-variables can be restricted directly (either individually or the sum of a set of K's), or indirectly by also implementing a maximum layover time. It is also possible to include the K-variables in the objective function.

Furthermore, constraint (6.34) for the realizations of the layover times at station n' would become:

$$\tilde{d}_{n'}^{t',z} \geq \tilde{a}_{n'}^{t,z-(k_{n'}^{t,t'}+K_{n'}^{t,t'})} + ml_{n'}^{t,t'}, \quad \text{for } z = k_{n'}^{t,t'} + 2, ..., Z. \tag{6.36}$$

However, this equation is non-linear, because it has the binary variable $K_{n}^{t,t'}$ in the index. Using a sufficiently large *big-M*, we can linearize the equation:

$$\begin{aligned} \tilde{d}_{n'}^{t',z} &\geq \tilde{a}_{n'}^{t,z-(k_{n'}^{t,t'}+1)} + ml_{n'}^{t,t'}, \\ \tilde{d}_{n'}^{t',z} &\geq \tilde{a}_{n'}^{t,z-k_{n'}^{t,t'}} + ml_{n'}^{t,t'} - K_{n'}^{t,t'} \cdot M, \end{aligned} \tag{6.37}$$

where M is large enough to make the latter constraint redundant in the case that $K_{n}^{t,t'} = 1$. In that case, the former constraint is binding. When $K_{n}^{t,t'} = 0$ the latter constraint is binding, and automatically implies that the first constraint is satisfied. Or in other words, when the departure of train t' is linked to the prior of the two possible arrivals of train t ($K_{n}^{t,t'} = 1$), then the first of the two equations tells that train t' cannot depart, unless the prior of the possible trains from line t has arrived at least $ml_{n'}^{t,t'}$ minutes ago. Because of the *big-M*, the second equation is redundant. When the departure of train t' is linked to the later of the two possible arrivals of train t ($K_{n}^{t,t'} = 0$), then the second of the two equations is not redundant anymore and tells train t' to wait with its departure to at least $ml_{n'}^{t,t'}$ minutes after the arrival of the later of the two trains from line t. The first equation also implies that the train has to wait for the earlier of the two possible trains, but this train is always earlier than the other one, since it is operated in an earlier cycle.

More variability in the number of cycle offsets can be modeled by increasing the number of K-like binary variables and additional, and more complicated *big-M* constraints.

6.7.2 Unknown Train Order

In the stochastic model as described so far, it is important to have a known sequence of events per infrastructure element. This enables us to formulate restrictions between events such as headways and passenger transfers. When the sequence of events is unknown, we can still model these restrictions, but binary variables have to be introduced as will be discussed below.

The order of the trains is mainly important for the constraints concerning the headways. When it is known that train t is earlier than train t', the departure headway constraint can be easily formulated linearly as

$$h \leq d_n^{t'} - d_n^t \leq T - h, \tag{6.38}$$

where $d_n^{t'}$ is the departure of train t' from station n. However, when the order of the trains is unknown a priori, this is too restrictive: it is incorrect when the train order is reversed. Therefore, we introduce a binary variable $O_{n,n'}^{t,t'}$ indicating the train order: $O_{n,n'}^{t,t'} = 0$ if train t is scheduled before train t' on the trip from station n to station n' and $O_{n,n'}^{t,t'} = 1$ if train t is scheduled after train t'. The headway restriction at departure then becomes:

$$h \leq d_n^{t'} - d_n^t + O_{n,n'}^{t,t'} \cdot T \leq T - h, \text{ for } t, t' : t \neq t' \in \theta_{n,n'}, \tag{6.39}$$

where $\theta_{n,n'}$ is the set of trips on the track from station n to station n'. The combination of these inequalities for t, t' and t', t makes sure that $O_{n,n'}^{t,t'} + O_{n,n'}^{t',t} = 1$. This also means that, if the trains can be ordered numerically, equation (6.39) only has to hold for $t, t' : t < t' \in \theta_{n,n'}$

To avoid the possibility of overtaking, an arrival headway constraint with the same binary order variable has to be included in the model as well. When overtaking at a certain station is not possible, the value of the order variables before and after this station have to be equal.

For the realizations, the same $O_{n,n'}^{t,t'}$-variable can be used. Equation (6.40) states that train t' leaves station n at least h later than train t. Of course, this should only hold when train t' is scheduled after train t. The term $O_{n,n'}^{t,t'} \cdot M$ makes this equation redundant when train t' is scheduled before train t:

$$h \leq \tilde{d}_n^{t',z} - \tilde{d}_n^{t,z} + O_{n,n'}^{t,t'} \cdot M, \text{ for } z = 1, ..., Z, \text{ and } t, t' : t \neq t' \in \theta_{n,n'}. \tag{6.40}$$

Any pair of events of which the order is unknown beforehand, and for which constraints have to be formulated, needs such a binary order variable $O^{t,t'}_{n,n'}$. This means that many binary variables are introduced into the model, which complicates the solvability.

Other constraints that need to be formulated concern the succession of the cycles. In other words, all trips in set $\theta_{n,n'}$ in cycle $z-1$ have to depart from station n before any trip in $\theta_{n,n'}$ in cycle z can depart from that station. In fact, a minimal headway h has to be respected:

$$h \leq \tilde{d}^{t',z}_n - \tilde{d}^{t,z-1}_n, \text{ for } z = 2, ..., Z, \text{ and } t, t' : t \neq t' \in \theta_{n,n'}. \tag{6.41}$$

However, this leads to $|\theta_{n,n'}|^2$ restrictions for each track and each realization. This number can be reduced by introducing a fictitious last train, t^* in $\theta_{n,n'}$. This means that we need $|\theta_{n,n'}|$ new constraints, ensuring that train t^* in cycle $z-1$ does not depart from station n before all other trains in $\theta_{n,n'}$ in cycle $z-1$ have left station n. Another $|\theta_{n,n'}|$ constraints make sure that no train in cycle z departs from station n before train t^* in cycle $z-1$ has departed from station n. Note that there is no minimal headway needed for the second set of constraints.

$$\begin{aligned} h &\leq \tilde{d}^{t^*,z-1}_n - \tilde{d}^{t,z-1}_n, && \text{for } z = 2, ..., Z, \text{ and } t \in \theta_{n,n'}, \\ 0 &\leq \tilde{d}^{t,z}_n - \tilde{d}^{t^*,z-1}_n, && \text{for } z = 2, ..., Z, \text{ and } t \in \theta_{n,n'}. \end{aligned} \tag{6.42}$$

By introducing the fictitious train t^*, the number of constraints for the cycle succession decreases from $|\theta_{n,n'}|^2$ to $2 \cdot |\theta_{n,n'}|$. Train t^* is not considered to be an element of $\theta_{n,n'}$.

6.7.3 Modeling the Timetable within the Cycle Interval

Direct application of the cycle offsets as presented in Section 6.7.1 requires careful calculation of running times and layovers between trains. Before the construction of the model, it has to be clear where those offsets are needed and which values are possible for the corresponding integer variables.

Therefore, a new formulation of the model has been developed. It is closer to the PESP-model, because all arrival and departure times attain values between 0 and the cycle time T. This new model has the advantage that it does not need any network investigation before it can be implemented. The disadvantage of this new representation is the increase in the number of binary variables.

In order to model the occurrence of all events within the $[0...T)$ interval, a binary variable C is introduced for each process that has to be scheduled. For example,

$Cr^t_{n,n'}$ indicates whether the running time of train t from station n to station n' passes the cycle bound ($Cr^t_{n,n'} = 1$) or not ($Cr^t_{n,n'} = 0$). An important assumption for the new model is that none of the processes has a planning time larger than or equal to T. But in fact, this can always be avoided by introducing dummy events, which divide a long process into two or more shorter processes.

Now the running time of train t from station n to station n', with minimal running time $mr^t_{n,n'}$, can be modeled by the following equation:

$$mr^t_{n,n'} + \sigma r^t_{n,n'} = a^t_{n'} - d^t_n + Cr^t_{n,n'} \cdot T. \tag{6.43}$$

Similar equations apply to other process types, such as dwell times and headways. Note that it is almost always possible to set the C-variable to 1. This may lead to excessive amounts of supplements in the timetable. To prevent this from happening, either the sum of sets of C-variables can be constrained, or the (sum of) individual supplements can be bounded.

The equations that prohibit earliness are the same as in equation (6.22):

$$\begin{aligned} \tilde{d}^{t,z}_n &\geq d^t_n + z \cdot T, \quad \text{for } z = 1, ..., Z, \\ \tilde{a}^{t,z}_n &\geq a^t_n + z \cdot T, \quad \text{for } z = 1, ..., Z. \end{aligned} \tag{6.44}$$

However, the binary variables imply that big-M equations have to be formulated for each realization of a process with a binary variable. An example is given for the running time of train t from station n to station n'.

$$\begin{aligned} \tilde{a}^{t,z}_{n'} &\geq \tilde{d}^{t,z}_n + mr^t_{n,n'} + \delta r^{t,z}_{n,n'} - Cr^t_{n,n'} \cdot M, \quad && \text{for } z = 1, ..., Z, \\ \tilde{a}^{t,z}_{n'} &\geq \tilde{d}^{t,z-1}_n + mr^t_{n,n'} + \delta r^{t,z-1}_{n,n'}, && \text{for } z = 1, ..., Z. \end{aligned} \tag{6.45}$$

When the arrival is planned in the same cycle as the departure, $Cr^t_{n,n'} = 0$ and the first equation is binding. When this first equation is satisfied, the second one is satisfied as well. In the case that the arrival is one cycle later than the departure, $Cr^t_{n,n'} = 1$ and the first equation becomes redundant. Now the second equation is binding.

Combination of unknown train order and $[0, T)$-modeling The combination of modeling within the cycle interval $[0, T)$ described in this section, and unknown train orders as described in Section 6.7.2 leads to some additional modeling features.

First of all, it has to be noted that an arrival time can be in a later period than the departure of the same trip. This means that the order of the departures within the cycle and the order of the arrivals within the cycle for a pair of trains are not necessarily the same. This implies that separate order variables are needed for the departures and arrivals of each pair of trips on the same track. For trains t and t' from station n to n' we call them $Od_n^{t,t'}$ and $Oa_{n'}^{t,t'}$ for departure and arrival, respectively.

The separation of the departure and arrival orders may lead to illegal overtakings. This can be avoided by introducing new equalities of binary order variables and cycle offset variables. Then for each pair of non-overtaking trains, the following restriction always holds:

$$Oa_{n'}^{t,t'} = Od_n^{t,t'} - Cr_{n,n'}^t + Cr_{n,n'}^{t'}. \tag{6.46}$$

The equality states that the train order variable at arrival is the same as at departure when neither or both trains pass the cycle bound. The train order within the cycle is reversed when exactly one of the two trains passes the cycle bound. This can be seen as follows: for the trains t and t' from station n to station n' the departure times d_n^t and $d_n^{t'}$, and the arrival times $a_{n'}^t$ and $a_{n'}^{t'}$ have to be determined within the $[0,T)$ interval. The planned running times of trains t and t' are $r_{n,n'}^t$ and $r_{n,n'}^{t'}$. Now define the running time constraints

$$\begin{aligned}\hat{a}_{n'}^t &= d_n^t + r_{n,n'}^t,\\ \hat{a}_{n'}^{t'} &= d_n^{t'} + r_{n,n'}^{t'}.\end{aligned} \tag{6.47}$$

When $r_{n,n'}^t$ does not cross the cycle bound $Cr_{n,n'}^t = 0$ and $a_{n'}^t = \hat{a}_{n'}^t$. Otherwise $Cr_{n,n'}^t = 1$ and $a_{n'}^t = \hat{a}_{n'}^t - T$; in other words $a_{n'}^t = \hat{a}_{n'}^t - Cr_{n,n'}^t \cdot T$. And similarly $a_{n'}^{t'} = \hat{a}_{n'}^{t'} - Cr_{n,n'}^{t'} \cdot T$.

Subtracting these two equations and dividing by the cycle time leads to

$$\frac{a_{n'}^t - a_{n'}^{t'}}{T} = \frac{\hat{a}_{n'}^t - \hat{a}_{n'}^{t'}}{T} - Cr_{n,n'}^t + Cr_{n,n'}^{t'}. \tag{6.48}$$

It is easy to verify that the binary order variables have been defined such that there is an $x \in [0,T)$ for which $Oa_{n'}^{t,t'} = (a_{n'}^t - a_{n'}^{t'} + x)/T$ and $O\hat{a}_{n'}^{t,t'} = (\hat{a}_{n'}^t - \hat{a}_{n'}^{t'} + x)/T$. Adding x/T to both the left and right hand side of equation (6.48) gives

$$Oa_{n'}^{t,t'} = O\hat{a}_{n'}^{t,t'} - Cr_{n,n'}^t + Cr_{n,n'}^{t'}. \tag{6.49}$$

It is obvious that the order of the trains remains the same along the trip from station n to station n' when no overtaking takes place and the cycle offsets are not used: $O\hat{a}_{n'}^{t,t'} = Od_n^{t,t'}$. Inserting this into equation (6.49), confirms that equation (6.46) is correct.

Reducing the number of binary C-variables The solvability of the problem is usually negatively correlated to the number of binary variables. Therefore, it is useful to look for possibilities to decrease the number of cycle offset variables.

Above, the model is described in such a manner that all processes require a binary cycle offset variable, because each process (running time, dwell time, headway, connection, etc) may cross the cycle bound. When some of the processes do not have a cycle offset variable, it is not possible to plan this process such that it starts before the cycle bound and ends after the cycle bound. However, when we do not assign these cycle offset variables to dwell times of which we know that they are never planned longer than ε minutes, and we allow to plan everything in a $[-\varepsilon, T+\varepsilon]$ interval, then we hardly affect the model. The headway inequalities will make sure that all planning times on the same infrastructure are still within the cycle time T. Then one can delete all cycle-offset variables associated with processes that have a maximal process time less than ε. A final prerequisite is that for all circular relations in the model, which relate trains of the same line, but z cycles apart, at least z processes need a cycle offset variable.

6.8 Allowing Homogenization

Chapter 5 describes how homogenization of the timetable can contribute to a higher reliability of the timetable. In this section, we first explain how homogenization can be modeled within our stochastic timetabling model. Thereafter, we again investigate the The Hague/Rotterdam – Utrecht case to analyze how the stops have to be assigned to the train lines to obtain a minimal average arrival delay. We still assume in this section that the line endpoints and routes are known.

6.8.1 Modeling Stops and Running Times

Two additional features have to be modeled to allow for homogenization in our stochastic model. Above all we need variables to indicate which trains stop at a certain station, and which trains do not. Furthermore, the running times have to be adjusted to the dwell pattern: when a train stops somewhere, it needs additional deceleration and acceleration time.

For each possible stop a binary variable is introduced: S_n^t equals 1 if train t stops in station n, and 0 if not. The new minimal running times can be calculated with these binary variables. However, first three minimal process times have to be defined for each trip from n to n'. The first one is the *minimal running time on the fly*: $mrf_{n,n'}^t$. This is the minimal running time for trains which pass both stations n and

n' at the maximal allowed speed. Furthermore we define the acceleration loss mra^t_n, which is the additional time that is needed when a train stops at station n at the beginning of the trip. Likewise, the deceleration loss $mrd^t_{n'}$ is needed for a train that is dwelling at station n' at the end of the trip. This leads to the following minimal running time:

$$mr^t_{n,n'} = mrf^t_{n,n'} + S^t_n \cdot mra^t_n + S^t_{n'} \cdot mrd^t_{n'} \tag{6.50}$$

Note that in practice trains always dwell at all begin and endpoints of a line. This implies that $S^t_n = 1$ for both line endpoints n of train t. Furthermore, when station n is not a line endpoint, S^t_n applies both to the deceleration loss preceding station n and the acceleration loss following station n.

Furthermore, the minimal stopping time ms^t_n of train t at station n can be defined as:

$$ms^t_n \geq S^t_n \cdot msp^t_n, \tag{6.51}$$

where msp^t_n is the provisional minimal stopping time in case that train t stops at station n.

Now it is easy to implement restrictions on the binary dwelling variables S^t_n. For example, for a certain set of trains Θ^i, Q^i_n trains have to dwell at station n:

$$\sum_{t\in\Theta^i} S^t_n = Q^i_n, \text{ for } i = 1, ..., I, \tag{6.52}$$

where I is the number of restrictions of this kind.

Besides the use of these optional stops for the homogenization of the train services, the optional stops can also be used to enable a cargo train to enter a siding to be passed by a faster train. In case of optional overtaking, freedom is needed in the train order, too (see Section 6.7.2).

6.8.2 Case The Hague/Rotterdam – Utrecht

The practical homogenization case presented in Section 5.4 is used here again. However, here we use our stochastic model including its homogenization possibilities described above. This means that we do not only compare two predetermined situations, but we are able to redistribute the train stops optimally with respect to the given disturbance distribution. On top of that, we can also optimally allocate the available buffers and supplements.

The case consists of 8 trains per direction per 30 minutes cycle. These lines are shown in Figure 6.5, but the stations where the trains dwell are unknown beforehand. These eight lines are split into trips which always start and end at two neighboring stations. This leads to 56 trips per direction. Not counting the line endpoints, where trains always dwell, there are 48 line–station combinations for which a binary variable has to be introduced. Note that the same binary S_n^t-variables are used for both directions of the same train line.

The total number of dwells per station per 30 minutes per direction is fixed, and is the same as in the 2004 peak hour timetable of NS. Practical circumstances decrease the number of binary variables a little bit. All four trains dwell at Rotterdam Alexander (Rta), lines 9700 and 9800, which end in Gouda Goverwelle, have to dwell in Gouda, and the only dwell in Vleuten is assigned to the 8800-line. This reduces the total number of binary variables to 36. Additional restrictions prescribe that exactly three of the four lines from Rotterdam Central dwell in Gouda, and two of the three lines from The Hague dwell in Gouda. Furthermore, one of the three lines from Rotterdam Central dwells in Woerden, and one of the two lines from The Hague does the same.

The maximal amount of running time supplement is limited per train to 7.92% of the minimal running time, which is the same as in the Haarlem–Maastricht case of Sections 4.4 and 6.5. There are no dwell time supplements.

For the comparisons, we also evaluated both dwell patterns from Section 5.4. Note that the supplements and buffers are also optimized for these two dwell patterns. With respect to this heterogeneous situation, the overtaking in Woerden (in western direction the 14000 line is overtaken by the 500 and the 20500 lines) is discarded and the local train 14000 leaves Utrecht after the intercity trains 500 and 20500, leading to shorter stops in Woerden and Gouda. This is the same train order as for the homogeneous situation from the same section and the train order used for the optimizations in this section.

Results The optimization of the dwell pattern, supplements and buffers on the The Hague/Rotterdam–Utrecht corridor has been performed for different disturbance distributions. Both running time disturbances and dwell time disturbances are applied. In case of running time disturbances, it is known beforehand where the disturbances occur. However, in case of dwell time disturbances, the location of the disturbances depends on the *choice* for dwell locations. All disturbances in this section are absolute disturbances, independent of minimal dwell or running times. Relative running time disturbances have not been analyzed, because there are three or four stations at all three sections (The Hague–Gouda, Rotterdam–Gouda, and Gouda–Utrecht) of 20 minutes in this case. Relative running time disturbances would probably lead to similar results as the absolute running time disturbances.

First 50 realizations are used for each disturbance distribution to optimize the dwell pattern, supplements and buffers. Then the dwell pattern is fixed and the supplements and buffers are re-optimized for 200 realizations from the same distribution.

The experiments with running time supplements show other results than the experiments with dwell time disturbances, especially with respect to the optimized dwell pattern. For cases with running time disturbances, we see that the optimal timetable has more homogeneous trains. For different disturbance distributions and levels, we find slightly different dwell patterns. However, in all cases, we find that the long distance trains dwell at more stations, and the short distance trains skip some stations. Despite the varying results, we find dwell patterns similar to the homogeneous situation in Figure 6.5(b). This is the dwell pattern used for the homogeneous situation in Section 5.4.

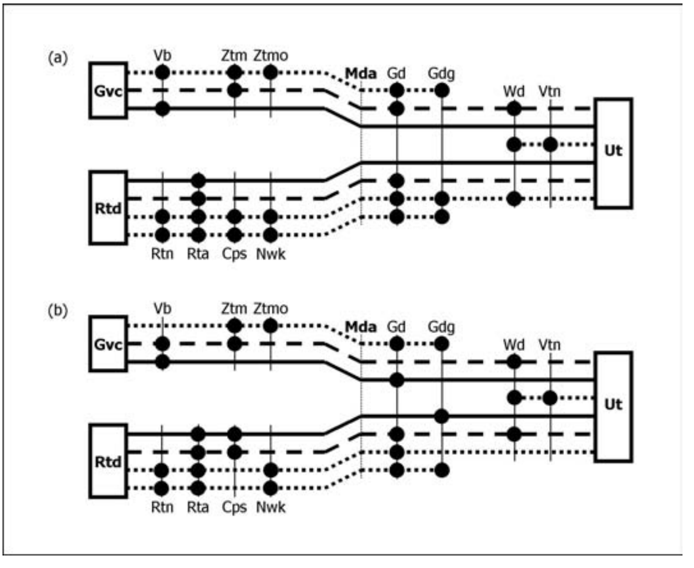


Figure 6.5: Heterogeneous (a) and homogeneous (b) dwell patterns from Section 5.4.

The experiments with dwell time disturbances show different results. The dwell patterns remain more heterogeneous in the optimal situation, although most results show a few stops shifting from short distance trains to long distance trains. Two possible explanations are given here. First, it is important to recognize that trains with many stops incur more delays. To reduce the delay propagation throughout the network it is beneficial to incur delays on short trains. In our case, these are the 9700 and 9800 short distance trains from Rotterdam and The Hague to Gouda Goverwelle. Indeed, it is observed that the 14000-line Rotterdam–Utrecht is more homogenized than the 9700 Rotterdam–Gouda Goverwelle. The short 8800-line Woerden–Utrecht should dwell everywhere in all cases. A second reason is the spacing of the trains over the 30-minute cycle. The order of the trains between Woerden and Utrecht has a large influence on the time distances between the trains on the Rotterdam–Gouda and The Hague–Gouda branches. In fact, these time distances are such that there is much time for additional stops on the 9700 line and especially on the 9800 line. This can probably be contributed to the fact that the used train order is the same as in the current heterogeneous situation.

Table 6.8 shows the results. The heterogeneous and homogeneous situations refer to the dwell patterns in Section 5.4, which are repeated in Figure 6.5. However, the supplements and buffers are optimally reallocated by the stochastic model for the different disturbance distributions. The average delay reductions are with respect to the re-optimized heterogenous situation. Although the dwell patterns of the heterogeneous and homogeneous timetables are fixed, the supplements and buffers differ between the different disturbance experiments. The optimal timetables are also different for all experiments; they even have different dwell patterns.

	Dwell Time Disturbance		Running Time Disturbance		Average Delay (in minutes)				
Experiment	Probability	Average size	Probability	Average size	Heterogeneous	Homogeneous	Reduction	Optimum	Reduction
1	100%	0.8	-	-	1.43	0.90	36.8%	0.88	38.2%
2	100%	1	-	-	2.10	1.33	36.5%	1.29	38.6%
3	-	-	100%	0.6	1.38	1.06	23.0%	1.06	23.0%
4	-	-	100%	0.8	2.33	1.85	20.5%	1.79	29.8%
5	20%	0.8	80%	0.8	2.20	1.68	23.6%	1.64	25.6%
6	50%	0.8	50%	0.8	1.99	1.44	27.7%	1.42	28.5%
7	80%	0.8	20%	0.8	1.68	1.15	31.4%	1.13	32.8%

Table 6.8: Results for the dwell pattern optimization.

Overall, the results show that homogenization is beneficial. And although the optimal dwell pattern differs slightly for most experiments, they all have one thing in common: the fast trains become slower and the slow trains become faster. This is especially true for the more crowded Rotterdam–Gouda branch. In all cases, the homogeneous situation chosen in Section 5.4 produces results close to the optimal timetable and much better than the heterogeneous situation. The optimal dwell patterns for experiments 2 and 3 are shown in Figure 6.6.

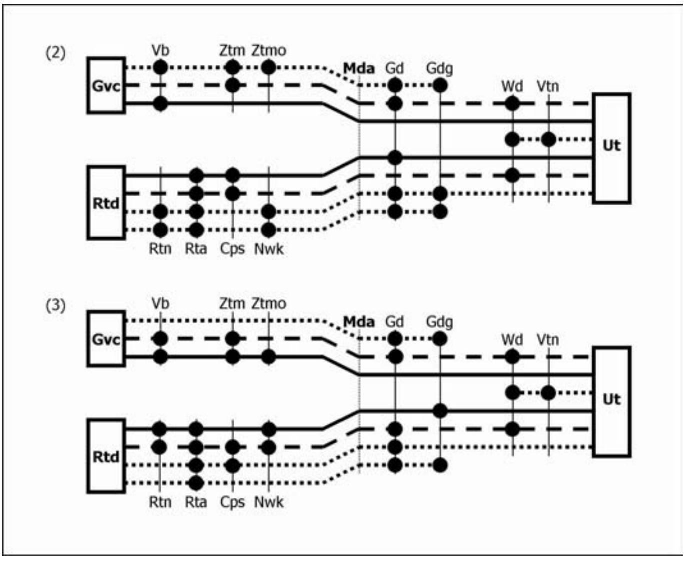


Figure 6.6: Optimal dwell patterns for experiment (2) with dwell time disturbances, and experiment (3) with running time disturbances.

For cases with only running time disturbances, homogenization is much more helpful than for cases with only dwell time disturbances. The results also show a positive correlation between the extent of the homogenization and the delay decrease.

The apparent differences in the results for different disturbance distributions asks for a detailed investigation of real world disturbances.

The experiments with only running time disturbances are solved within a few minutes. When dwell time disturbances are involved the model seems to be harder to solve. This is probably due to the fact that it is unknown a priori where the disturbances are incurred, because it is unknown which lines dwell at which stations.

SSHR and SAHR In Chapter 5, two heterogeneity measures were developed to forecast the reliability of a timetable. There the conclusion was that a high value of the Sum of Shortest Headway Reciprocals (SSHR) or the Sum of Arrival Headway Reciprocals (SAHR) indicates a poor reliability for the considered timetable. Smaller values predict a better on-time performance.

In Section 5.4, a heterogeneous timetable (see Figure 5.9) and a homogeneous timetable (see Figure 5.10) were evaluated. The heterogeneous and homogeneous timetables in this section are similar, but not exactly the same. The dwell patterns of both timetables from Section 5.4 (see Figure 6.5) remain the same. However, these timetables are re-optimized with the stochastic model by adjusting supplements and buffers. This leads to different heterogeneous and different homogeneous timetables for all experiments, all with different SSHRs and SAHRs. Furthermore, the case in Chapter 5 shows large values for the SSHR and SAHR in the heterogeneous situation, because of the overtaking of the 14000 local train from Utrecht to Rotterdam in Woerden. This overtaking is not present in the timetables in this chapter.

In Table 6.9, the SSHRs and SAHRs for the six trajectories of experiments 2 and 3 are presented for the heterogeneous, homogeneous and optimal timetables.

SSHR							SAHR					
experiment 2			experiment 3				experiment 2			experiment 3		
heterogeneous	homogeneous	optimal	heterogeneous	homogeneous	optimal	trajectory	heterogeneous	homogeneous	optimal	heterogeneous	homogeneous	optimal
0.77	0.67	0.74	0.78	0.71	0.86	Gvc-Gd	0.61	0.64	0.60	0.64	0.68	0.71
3.21	1.41	1.42	3.21	1.39	1.88	Rtd-Gd	1.65	1.09	1.08	1.65	1.18	1.30
4.62	3.33	3.54	4.11	3.39	3.44	Gd-Ut	3.29	2.89	2.83	3.28	2.97	2.99
3.66	2.97	2.75	3.50	2.90	2.92	Ut-Gd	2.22	2.14	2.06	2.28	2.11	2.17
0.90	0.80	0.89	0.79	0.69	0.87	Gd-Gvc	0.61	0.62	0.61	0.61	0.62	0.62
3.23	1.36	1.34	2.91	1.37	1.82	Gd-Rtd	2.34	1.12	1.10	2.33	1.18	1.49

Table 6.9: The SSHR and SAHR for scenario 2 with dwell time disturbances, and scenario 3 with running time disturbances.

The values of the SSHR and SAHR for the heterogeneous and homogeneous dwell patterns are relatively close to the values found in Section 5.4, both for scenario 2 and 3. The SSHR and SAHR for Utrecht-Gouda in the heterogeneous situation of this earlier section should not be compared to the SSHR and SAHR of the heterogeneous situation in this section. That is because the overtaking in Woerden in Section 5.4 does not take place in this section.

For the most utilized parts of the infrastructure, Rotterdam-Gouda and Gouda-Utrecht and vice versa, it is found that the SSHR and SAHR for the optimal situation are comparable to the values for the homogeneous situation.

On the less utilized tracks from The Hague to Gouda and vice versa, the SAHR for the optimum is comparable to the values for the heterogeneous and homogeneous situation. The SSHR-values are even higher for the optimal timetable. This can be explained by the fact that the timetable characteristics for the most crowded parts of the infrastructure are more important for the whole network. That is due to the fact that trains are closer together on crowded tracks, and more interaction between the trains is expected there.

Furthermore, the SSHR and SAHR seem to be more important for the scenario with dwell time disturbances. This is understandable, since it is advantageous to have equally disturbed trains. In case of dwell time disturbances, this means that all trains should have an equal number of stops. This leads to smaller values of the SSHR and SAHR.

6.9 Discussion and Conclusions

The stochastic timetabling model presented in this chapter is able to provide an optimal timetable with respect to the objective and a predetermined set of disturbances. The model is a large scale linear program which integrates the timetabling problem and the evaluation of this timetable for multiple realizations.

The main objective is to construct a reliable timetable with minimal average delay. However, other aspects of the timetable, such as running time supplements or passenger transfer times, can be included in the objective function as well.

The decision variables in the model are the departure and arrival times. These departure and arrival times determine the available supplements and buffers in the timetable, and also the transfer times for passengers and the layover times for the rolling stock. The supplements and buffers directly influence the reliability of the model. The planned supplements and the passenger transfer times determine the planned travel times for passengers. The presented model is, to our knowledge, the first timetable optimization model which directly takes delay propagation into account.

The model is very useful in developing reliable new timetables and improving existing timetables on a subnetwork. Our computational results in Section 6.5.2 showed that, within the model, delay reductions of over 30% are possible.

The disturbances applied in the experiments are not the real life disturbances. However, the sensitivity analysis shows that the optimal timetable with respect to one disturbance distribution also performs much better than the original timetable for a wide range of other disturbance distributions.

Some of the other advantages of the model are listed here. Listing these advantages shows that the model is flexible in many ways and useful for analyzing many timetabling problems.

- Most timetabling restrictions can be incorporated in the model. This is necessary to be sure that the optimal timetable provided by the model is feasible.

- Because it is an optimization model instead of an evaluation model, no trial-and-error procedure is needed to find an optimal timetable.
- Any kind of probability density function can be applied for the disturbances in the model.
- Besides the average delay, other timetable characteristics such as supplements can be included in the objective function. It is even possible to maximize punctuality by using binary variables. All of these characteristics can be weighed (for the number of passengers) in the objective function.
- The model is very flexible. Depending on the problem at hand, one of the discussed model variants can be used. Several extensions have been discussed already in Sections 6.6 through 6.8, but further variants of the model certainly exist.
- By fixing all departure and arrival times, the model can also be used to evaluate known timetables.
- The model is explained as a large scale linear program. This does usually not lead to an integer timetable. The model can easily be rewritten as a Mixed Integer Program, where departure are arrival times are integer. This leads to a more complicated optimization and less timetabling possibilities. Usually, the average delay will go up.
- The choice whether to stop at a station or not provides much flexibility to the model. This option can be used to optimally homogenize a timetable, but also for determining the optimal location for possible overtakings of cargo trains. Binary order variables are needed as well in that case.

On the other hand, we see the following drawbacks of the model.

- The timetable which is found, is optimal with respect to the predetermined set of disturbances. A different set of predetermined disturbances will lead to a different timetable. Although the solutions seem to perform reasonably well under different sets of disturbances from both the same and other distributions, it is important that the set of disturbances is a good representation of real-world disturbances.

  However, in Section 6.3 it is shown for the case with two trajectories and a given disturbance distribution, that the solution converges to the optimum when the number of realizations goes to infinity. Although not proven, this result can intuitively be extended to the complete model.

- It is not easy to assess the variability of the results with respect to the disturbances. Therefore, it is hard to predict how many realizations are required to obtain a certain reliability of the model results. Large numbers of realizations can lead to enormous models (number of variables and constraints) and excessive running times.

- Most of the extensions of the model lead to binary variables. These binary variables make the problem much harder to solve. The complexity seems to increase even more when delays depend on the values of binary variables. This may be the case for binary variables indicating whether a train dwells at a certain station in combination with dwell time disturbances.

- It is possible to model an operational decision rule, such as the waiting time rule discussed at the end of Section 6.6.3. However, many binary variables and *big-M* constraints are needed to avoid unnecessary delays. This means that the resulting problem formulation will be hard to solve.

Despite the disadvantages mentioned here, the presented model is very well capable of representing a real life timetable and its operation. Applying the model can lead to sizable reliability improvements. However, any optimization model depends on good input data. Hence it is very important to get good estimates of the disturbances. Apart from reliability, the model can optimize other timetable qualities as well.

Chapter 7

Discussion and Conclusions

An increasing demand for railway traffic asks for a better utilization of the available infrastructure. Structural changes in planning and operating the railway system are necessary to enable this without excessive delays. New methods are needed to this end, and that was the aim of this thesis. Chapters 4 through 6 present innovative methods for the improvement of timetables. The main results from these chapters are summarized in Section 7.1 to put them in a broader perspective. These results are evaluated in Section 7.2 with respect to the research steps described in Section 1.3.

Despite the fact that several questions are answered in this thesis, many others are still open. Furthermore, new questions were raised while answering others. Some of these questions are posed in Section 7.3 and are challenging subjects for further research.

7.1 Results of this Thesis

This section discusses the results of Chapter 4 on supplement allocation, Chapter 5 on heterogeneity, and Chapter 6 on stochastic timetable optimization.

7.1.1 Results on Running Time Supplements

In Chapter 4, a numerical model is introduced, which enables the optimization of the supplement allocation for a single train. The optimization is with respect to expected average arrival delays at the intermediate stations and the end station of a train, given a certain disturbance distribution per trip.

Present day timetabling at NS aims at a proportional allocation of the supplements. However, the results from the model show that this proportional allocation is not optimal with respect to the objective. When the total amount of supplements equals the expected total amount of disturbances, and each trip is equally and exponentially disturbed, then the allocation as depicted in Figure 4.6 is found. This allocation proposes that less than average supplement should be allocated to the first and last trips, regardless of the number of trips of the line. This principle can be applied to all lines of a whole network. So with relatively little adaptations to the complete timetable of NS, substantial improvements in the performance of NS can be achieved. Compared to the investments required in the physical infrastructure or rolling stock to achieve such a result, our approach yields considerable improvements against a fraction of the costs.

Although the result may be quite surprising, they can be explained quite intuitively. First of all, a delay reduction on one of the first trips will be measured at many stations ahead. Later reductions will only be measured at the last stations. Supplements at earlier trips are therefore more useful. On the other hand, we do not want too much supplements on the earlier trips, because supplements can not be used when the train has not incurred any disturbances yet. This leads to above average supplements on the middle part of the line. However, the supplements are, on average, slightly shifted to the front of the line.

As opposed to the shape of the supplement allocation, the size of the delay reduction does depend on the number of trips. When the number of trips increases, the average delay reduction also goes up. In fact, a decrease of 16% of the average arrival delays is reached for 10 trips. Given the effort that NS has put into other activities to reduce delays, this result can be classified as spectacular.

Section 4.5.2 discusses some practical implications of the model on *rounding*, *running time differences*, and *symmetry*.

7.1.2 Results on Heterogeneity

Chapter 5 discusses the heterogeneity of railway traffic. A heterogeneous timetable is a timetable with large scheduled speed differences between trains on the same track. Homogenization means that these speed differences are reduced. It is argued that homogenization leads to larger headways, and consequently, to less delay propagation, and hence more robust timetables.

To support this argument, two new timetable measures were developed: the Sum of Shortest Headway Reciprocals (SSHR), and the Sum of Arrival Headway Reciprocals (SAHR). These measures are not only able to measure the heterogeneity of a certain track, but they are also able to give a prediction of the reliability. Both measures have small values for homogeneous timetables with a high expected punctuality, and large values for heterogeneous timetables with a poor expected performance.

Both a practical and a theoretical case were analyzed. Both cases showed that the homogeneous timetable had much smaller delays than the original heterogeneous timetable. Moreover, this delay reduction was predicted by our new measures, the SSHR and SAHR.

The SAHR stresses that arrival headways are more important than departure headways. The SSHR is able to capture the headways and heterogeneity for an entire trajectory, as opposed to known headway measures, which focus on a single point of the infrastructure. To attain a more reliable timetable, it should be constructed such that both the SSHR and the SAHR are small. Although the measures can not be applied to a network as a whole, they can be applied to all individual trajectories separately.

Homogenization of the timetable does not only influence the reliability of the timetable. It also affects other timetable characteristics. In Section 5.5 some additional consequences of homogenization are discussed.

7.1.3 Results on Stochastic Timetable Optimization

Chapter 6 does not focus on a single timetable characteristic. In fact, the stochastic optimization model presented there does not use timetable measures or characteristics, but it directly relates a timetable to its on-time performance. Moreover, it does not only evaluate a timetable, but it optimizes the timetable with respect to a given objective function.

Again, on-time performance is the main objective considered, but other characteristics, such as travel time, can also be included in the model. The optimization of the model is with respect to a given set of disturbances. This set is a sample from a certain disturbance distribution, and is determined before the model optimization is started.

As an example, delays can be weighed against supplements to determine the optimal amount of supplements with respect to the disturbance distributions.

The stochastic timetable optimization model is a large linear program. In fact, it is related to recourse models (Klein Haneveld and Van der Vlerk, 2004). The first part of the constraints model the timetable restrictions. The second, much larger part of the constraints evaluates the on-time performance of this timetable under construction, given certain disturbances. These two parts are integrated, such that the evaluation part guides the timetabling part towards the optimal timetable.

Realistic situations can be modeled and optimized in reasonable time. The Haarlem–Maastricht/Heerlen case was solved in about 45 minutes. The model is very useful in developing reliable new timetables and improving existing timetables on a subnetwork. Our computational experiments with the stochastic model in Section 6.5.2 showed that delay reductions of over 30% are possible in the Haarlem–Maastricht case, which again can be classified as striking. The implementation of additional dependencies in the timetable, such as layovers and crossings with other trains around nodes, can improve the accuracy of the model predictions.

To the best of our knowledge, this is the first timetabling model which takes delay propagation explicitly into account, and optimizes the timetable directly, without cumbersome trial-and-error procedures. Above all, we expect a punctuality increase of several percent-points when the model is integrally adopted for a nationwide timetable.

7.2 The Research Steps Evaluated

The central research issue of this thesis was described in Section 1.3. That section also gave five steps to guide the research towards tackling this central issue. These research steps are related to the results of this thesis in this section. Although the five research steps are made in the order that they were posed, the answers appeared in a different order in this thesis. In this section we first look into these five research steps, and conclude with a discussion on the central issue of our research project.

Describe determinants of punctuality The first step considered the influence of planning rules and other aspects on punctuality. Although it is not exhaustive, Section 2.6 gives a broad overview of planning principles influencing punctuality. However, only qualitative considerations are given there. It is apparent from this thesis, that the complexity of timetables, rolling stock circulations, and crew schedules, makes it difficult to quantify these determinants.

Estimate the impact of different aspects on punctuality The quantification of the described influences is exactly the second research step. This thesis shows that, with some effort, reasonable and quantifiable measures can be developed. We managed to quantify the heterogeneity and headway allocation for a trajectory with the SSHR and SAHR in Chapter 5. The next step is to find quantitative measures for networks.

Compare the quality of different timetables The next step concerns the development of models or tools which can be used to compare timetable scenarios based on stability and robustness. In Chapter 5, we did not only quantify the heterogeneity of the train services on a certain line with the heterogeneity measures, but we could also make a comparison between several timetables. Timetables with low values for the SSHR and SAHR are expected to perform better. Furthermore, the simulation tool SIMONE was used to compare timetable scenarios in Chapters 4 and 5. Additionally, timetables were also compared with the stochastic optimization model in Chapter 6.

Apply the theoretical models in practical cases In Chapter 4 a numerical model was developed to allocate the available running time supplements optimally. It was used to validate the appropriateness of the current day supplement allocation in the Netherlands. An optimization with this model shows that a better allocation of the supplements is possible. Moreover, these practical results are supported by the results in Sections 6.2 and 6.5.

The results of Chapter 5 and Section 6.8 provide another important result, which is directly related to practice. They show that the punctuality can be improved by homogenization of the present day heterogeneous timetables.

Improve the timetable The stochastic timetabling model presented in Chapter 6 is, to our knowledge, the first timetabling model that takes delay propagation into account explicitly. The model does not optimize a timetable regarding some timetable characteristic which is supposed to correlate to the robustness; it minimizes the arrival delay directly. Therewith it can be a very helpful tool for the improvement of timetables.

Summarizing the above, we clearly worked towards the central issue of this thesis:

> ***The central issue in this thesis is to develop rules and instruments for supporting the generation of more reliable timetables.***

The stochastic optimization model presented in Chapter 6 is an innovative timetabling model. It not only enables the construction of a *more* reliable timetable, it also enables the construction of the *most* reliable timetable with respect to the predetermined disturbances and objective. After some further improvements, the model certainly has a high potential in the near future.

7.3 Further Research

Sections 7.1 and 7.2 give a range of answers and solutions to questions and problems in the area of railway reliability. However, many more interesting questions with respect to improving railway timetables exist. The first five research ideas given below concern the stochastic timetabling model presented in Chapter 6. More general research ideas are presented thereafter.

Analyze the practical case for the stochastic timetabling model in more detail The presented Haarlem–Maastricht case has been analyzed with the stochastic timetabling model in Section 6.5. However, not all details were included in the model. For example, dwell time supplements and cargo trains were not included. Implications for real world changes in the timetable can be predicted with more certainty when the case is analyzed with a more detailed model.

Convergence of the stochastic timetabling model The proof of convergence of the stochastic timetabling model for the case with two trips is given in Section 6.3. An overall convergence has not been proven yet. Moreover, the speed of convergence is unknown. The necessary number of realizations for the stochastic model can be determined when the speed of convergence is known.

Analyze conceptual extensions to the stochastic timetabling model The applicability of the stochastic model for homogenization of timetables was analyzed in Section 6.8. The necessity of binary variables increased the computational complexity of the model considerably. A detailed analysis of this complexity can be very helpful to see which kind of problems can and can not be solved. To this respect, all other mentioned extensions with binary variables and *big-M* constraints should be considered.

Special-purpose solver The stochastic model in Chapter 6 is implemented in OPL Studio and solved with the solver CPLEX. However, because CPLEX is a general-purpose solver, it can neither take advantage of the special characteristics and structures of the timetabling problem at hand, nor of those of recourse models. A self-developed special-purpose solver could be very helpful in finding the optimal solutions faster. A smarter use of computer memory by this solver will also enable to solve larger problems.

Other applications of the stochastic timetabling model The stochastic timetabling model in Chapter 6 can be applied to a wide range of other planning problems. The model can not only be applied to transportation systems like airlines and shipping, but also to other systems operating under uncertainty, such as production lines with complicated sequence requirements. However, the system that has to be optimized must be of cyclic nature. In all of the alternative applications, some of the restrictions might be redundant, and new restrictions might be necessary.

Apart from a further development of the stochastic model of Chapter 6, several other research areas are described.

Disturbance distributions The timetable evaluations and optimizations in this thesis are all based on certain sizes and frequencies of disturbances. This means that the timetables are assessed, or optimized, with respect to these disturbances. However, these are theoretical disturbance distributions. Although they are expected to be reasonable, nobody knows the exact real world disturbances. Real world measurements are only available for the total delays. There is no information on the distinction between primary and secondary delays, and the possible use of supplements. This is really a white spot in on-time performance evaluations in railway research. It is essential to determine the size and frequency of real world disturbances. The timetables can then be optimized with respect to these real world disturbances, and supplements can be planned there where the disturbances occur most often.

Detailed analysis of large stations In this thesis the focus has been on delay propagation on a network level. A detailed analysis of train movements at and around stations has not taken place. Still, train routing through stations can have a considerable impact on delay propagation. A better understanding of this subject is very useful for the construction of reliable timetables.

Quantify the qualitative considerations A large number of possible influences of planning principles are discussed qualitatively in Section 2.6. Only the *supplement allocation* and *heterogeneity* have been discussed quantitatively in Chapters 4 and 5, respectively. More insight in the other planning characteristics mentioned in Section 2.6, such as fixed corridors and line lengths, could help planners to construct better timetables. Hendriks (2004) provides a good starting point for more in-depth research of norms and planning concepts. Furthermore, Ybema (2000) already focused on the consequences of different lengths of train lines.

Above, a few of many possible intriguing research questions about railway reliability in general, and the models presented in this thesis can be found. Despite these remaining questions, a substantial contribution to the field of railway research has been delivered in this thesis.

Appendix A

Glossary

A.1 Terminology

The most important terms and abbreviations are gathered in this appendix. Terms used in the explanations in *italics* are also included in this appendix themselves. The abbreviations for timetable points can be found in appendix A.3

ACTS: cargo railway operator.

Betuwe Route: new cargo line, connecting the port of Rotterdam with the German hinterland. To be completed in 2007.

connecting train: train which offers a connection from another, the feeder, train.

cyclic timetable: repetitive schedule; train arrivals and departures are the same every cycle time.

delay: lateness of an event, e.g. a departure or an arrival.

disturbance: an initial perturbation of the timetable.

dwell time: time between arrival and departure at a station to enable passengers to alight and board.

ERS (European Rail Services): international cargo railway operator.

feeder train: train which offers a connection to another, the connecting, train.

High Speed Alliance: Joint venture of NS and Air France-KLM which will operate the high speed line from Amsterdam to Paris.

homogenization: decreasing the speed differences of the railway traffic.

HSA: *High Speed Alliance*.

HSL-Zuid: (High Speed Line-South) Dutch part of the high speed line from Amsterdam towards Paris. To be completed in 2007.

Nederlandse Spoorwegen: Netherlands Railways; main passenger railway operator in the Netherlands. Also operating abroad.

NoordNed: passenger railway operator of several peripheral lines in the northern part of the Netherlands.

NS: *Nederlandse Spoorwegen*.

NS Reizigers: part of *NS* which plans and operates the passenger trains.

NSR: *NS Reizigers*.

OHT: *One-Hour Timetable*.

One-Hour Timetable: sixty-minute timetable which can be copied to create a *cyclic timetable* for a whole day.

perceived travel time loss: weighted average of planned supplements and expected delay for passengers.

Platform Occupation Chart: graphical representation of the planned dwell times at a station.

POC: *Platform Occupation Chart*.

primary delay: initial delay, not caused by the delay or cancelation of another train, but due to a *disturbance*.

punctuality: reliability measure, measuring the percentage of trains running within a certain time margin from its scheduled time.

Rail4Chem: international cargo railway operator.

Railion: international cargo railway operator.

reliability: the ability of the railway system to function as the timetable indicates.

robustness: the ability of the railway system to operate normally despite disturbing influences.

running time: time between departure at one station and arrival at the next station.

secondary delay: knock-on delay; caused by a delay or cancelation of one or more other trains.

SAHR: *Sum of Arrival Headway Reciprocals*.

SSHR: *Sum of Shortest Headway Reciprocals*.

Sum of Arrival Headway Reciprocals: indirect measure for the heterogeneity of train traffic.

Sum of Shortest Headway Reciprocals: measure for the heterogeneity of train traffic.

stability: the ability of the railway system to return to normal, starting from a disturbed situation.

Syntus: passenger railway operator of several peripheral lines in the eastern part of the Netherlands.

time-distance diagram: graphical representation of the timetable.

Train Team: the principle of keeping train driver and conductor together for the extent of their duty.

WAD: *Weighted Average Distance*.

Weighted Average Distance: measure which indicates the average location of supplements on a certain line.

A.2 Parameters, Variables and Indices

The symbols used for the parameters, variables, indices and sets in this thesis are given below. The abbreviations for timetable points can be found in Appendix A.3.

A	Set of arrival events.
$A_x(\sigma_1)$	$x = 1, 2, 3, 4$. The four areas resulting from the partitioning of the positive (δ_1, δ_2) quadrant for a given value of σ_1.
a_n^t	Planned arrival time of train t at timetable point n.
$\tilde{a}_n^{t,z}$	Realized arrival time of train t at timetable point n in realization z.
$\hat{a}_n^t$	Planned arrival time of train t at timetable point n in a cyclic timetable, not corrected for crossing the cycle bound.
b_i^-	Smallest planned buffer between train i and $i+1$ on a certain trajectory.
$Cr_{n,n'}^t$	Binary variable to indicate whether the running time of train t between stations n and n' in a periodic timetable crosses the cycle bound or not.
d_n^t	Planned departure time of train t from timetable point n.
$\tilde{d}_n^{t,z}$	Realized departure time of train t from timetable point n in realization z.
D	Set of departure events.
D_n^t	Clock set related to the departure time of train t from timetable point n.

h	Minimal headway; in Chapter 5 denoted by h^{min}.
h_i^-	Smallest planned headway between train i and $i+1$ on a certain trajectory.
h_i^A	Planned headway at arrival between train i and $i+1$.
I	Number of discretization intervals in the numerical model for the optimization of the supplement allocation.
J	Average planned travel time supplements.
J^w	Average perceived planned travel time supplements.
$k_n^{t,t'}$	Number of cycles elapsed between trains t' and t at station n.
$K_n^{t,t'}$	Binary variable to indicate an additional elapsed cycle between trains t' and t at station n.
l	Lower bound
$ml_n^{t,t'}$	Minimal layover time of train t on train t' at station n.
$Ml_n^{t,t'}$	Maximal layover time of train t on train t' at station n.
$mpt_n^{t,t'}$	Minimal passenger transfer time of train t on train t' at station n.
$Mpt_n^{t,t'}$	Maximal passenger transfer time of train t on train t' at station n.
$mr_{n,n'}^t$	Minimal running time of train t from station n to station n'.
mra_n^t	Minimal acceleration loss of train t when stopping at station n.
mrd_n^t	Minimal deceleration loss of train t when stopping at station n.
$mrf_{n,n'}^t$	Minimal running time on the fly of train t from station n to station n'.
ms_n^t	Minimal dwell time of train t at station n.
Ms_n^t	Maximal dwell time of train t at station n.
msp_n^t	Provisional minimal dwell time of train t at station n.
N	Number of stations or timetable points.
$O_{n,n'}^{t,t'}$	Binary variable to indicate the order of trains t and t' between timetable points n and n'.
$Oa_n^{t,t'}$	Binary variable to indicate the order of arrival of trains t and t' at timetable point n in a cyclic timetable.
$Od_n^{t,t'}$	Binary variable to indicate the order of departure of trains t and t' from timetable point n in a cyclic timetable.
p	Process, such as running along the track, dwelling, or a headway time.
P^t	Set of processes being part of train line t.
Q_n^i	Number of trains from set Θ^i that has to dwell at station n.

$r_{n,n'}^t$	Planned running time of train t from timetable point n to timetable point n'.
s_n^t	Planned dwell (or stopping) time of train t at station n.
S_n^t	Binary variable to indicate whether train t dwells at station n or not.
T	Cycle time of a periodic timetable.
u	Upper bound
V	Size of the discretization intervals in the numerical model for the optimization of the supplement allocation.
w_j	weight for arrival j relative to other arrivals.
$w_{pt},\ w_r,\ w_s$	Relative weights for the duration of different travel time components: passenger transfers, running times, and dwell times.
w_Δ	Weight that determines the importance of delays relative to travel time.
Z	Number of realizations for the stochastic timetable optimization.
α_a	Number of passengers for which process a is the last part of their journey.
β_b	Total number of passengers using process b in their journey.
$\gamma_{n,n'}$	Number of passengers entering the railway system at station n, and leaving the railway system at station n'.
$\delta r_{n,n'}^t$	Disturbance on the running time of train t between stations n and n'. Sometimes just denoted by $\delta_{n,n'}^t$.
δs_n^t	Disturbance on the dwell time of train t at station n.
Δ	Average delay.
Δ_n	Arrival delay at station n.
$\Delta a_n^{t,z}$	Arrival delay of train t at station n in realization z. Sometimes just denoted by Δ_n^z.
$\Delta d_n^{t,z}$	Departure delay of train t at station n in realization z.
ε	Slack variable or maximal deviation from a certain value or interval.
π_a^z	Binary variable indicating whether arrival a in cycle z is punctual or not.
Π	Punctuality.
$\rho_{n,n'}^{t,t'}$	Running time difference between train t and train t' between timetable points n and n'.
$\sigma pt_n^{t,t'}$	Supplement on the passenger transfer time from train t to train t' at station n.
$\sigma r_{n,n'}^t$	Running time supplement of train t on the running time from timetable point n to timetable point n'.
σs_n^t	Dwell time supplement of train t at timetable point n.

τ	Clock value.
$\theta_{n,n'}$	Set of all trips from timetable point n to timetable point n'.
Θ	Set of all train lines.
Θ^i	Subset of train lines.

A.3 Timetable Point Abbreviations

In this appendix all official abbreviations for timetable points (stations, junctions, and others) which are used in this thesis are listed. The indications (B), (D) and (F) behind the names, indicate that the station is in Belgium, Germany or France, respectively.

abbreviation	**timetable point**
Ac	Abcoude
Ah	Arnhem
Ahbf	Aachen Hbf (D)
Almb	Almere Buiten
Amf	Amersfoort
Amfs	Amersfoort Schothorst
Aml	Almelo
Amr	Alkmaar
Apd	Apeldoorn
Apn	Alphen a/d Rijn
Asa	Amsterdam Amstel
Asb	Amsterdam Bijlmer
Asd	Amsterdam Centraal
Asdm	Amsterdam Muiderpoort
Ass	Amsterdam Sloterdijk
Atw	Antwerpen Centraal (B)
Bd	Breda
Bde	Bunde
Berch	Antwerpen Berchem (B)
Bet	Best
Bh	Bad Bentheim (D)
Bk	Beek-Elsloo
Bkl	Breukelen
Bmr	Boxmeer
Br	Blerick
Brn	Baarn
Brusz	Brussel-Zuid (B)
Btl	Boxtel
Cl	Culemborg
Cps	Capelle Schollevaar
Ddr	Dordrecht
Dn	Deurne
Dtc	Doetinchem
Dussel	Düsseldorf Hbf (D)
Dv	Deventer
Dvd	Duivendrecht
Dz	Delfzijl
Ec	Echt
Ed	Ede-Wageningen
Ehb	Eindhoven Beukenlaan
Ehv	Eindhoven
Ekz	Enkhuizen
Emn	Emmen
Es	Enschede
Frank	Frankfurt (Main) Hbf (D)
G	Gronau (D)
Gd	Gouda
Gdg	Gouda Goverwelle
Gdm	Geldermalsen
Gln	Geleen Oost
Gn	Groningen
Gp	Geldrop

Gr	Gorinchem	Nwk	Nieuwerkerk a/d IJssel
Gv	Den Haag HS	Odz	Oldenzaal
Gvc	Den Haag Centraal	Paris	Paris Nord (F)
Hann	Hannover Hbf (D)	Rd	Roodeschool
Hb	Hoensbroek	Rhn	Rhenen
Hdr	Den Helder	Rm	Roermond
Hfd	Hoofddorp	Rsd	Roosendaal
Hgl	Hengelo	Rta	Rotterdam Alexander
Hld	Hoek van Holland Haven	Rtd	Rotterdam Centraal
Hlds	Hoek van Holland Strand	Rth	Rotterdam Hofplein
Hlgh	Harlingen Haven	Rtn	Rotterdam Noord
Hlm	Haarlem	Sbk	Spaubeek
Hlms	Haarlem Spaarnwoude	Sgn	Schagen
Hn	Hoorn	Shl	Schiphol
Hnk	Hoorn Kersenboogerd	Sk	Sneek
Hrl	Heerlen	Sn	Schinnen
Ht	’s-Hertogenbosch	Srn	Susteren
Htn	Houten	Std	Sittard
Htnc	Houten Castellum	Stola	Stolberg Altstadt (D)
Hze	Heeze	Stv	Stavoren
Kfh	Kijfhoek	Tb	Tilburg
Koln	Köln Hbf (D)	Tl	Tiel
Kpn	Kampen	Ut	Utrecht Centraal
Krd	Kerkrade	Utg	Uitgeest
Ledn	Leiden Centraal	Utl	Utrecht Lunetten
Leer	Leer (D)	Vb	Voorburg
Lls	Lelystad Centrum	Vdg	Vlaardingen Centrum
Luik	Liège-Guillemins (B)	Vg	Vught
Lut	Geleen Lutterade	Vl	Venlo
Lw	Leeuwarden	Vry	Venray
Mas	Maarssen	Vs	Vlissingen
Mda	Moordrecht aansluiting	Vtn	Vleuten
Mrb	Mariënberg	Wd	Woerden
Msw	Maassluis West	Ws	Winschoten
Mt	Maastricht	Wt	Weert
Mtr	Maastricht Randwyck	Ww	Winterswijk
Munst	Münster Hbf (D)	Zbm	Zaltbommel
Ndb	Naarden-Bussum	Zl	Zwolle
Nh	Nuth	Zp	Zutphen
Nm	Nijmegen	Zsh	Zoetermeer Stadhuis
Nsch	Nieuweschans	Ztm	Zoetermeer

Ztmd	Zoetermeer Dorp	Zvt	Zandvoort
Ztmo	Zoetermeer Oost		

A.4 Train Line Numbering

The next table shows the 2004 train line numbering of passenger trains in the Netherlands. For almost all lines, one hundred consecutive numbers are reserved. For example the 500-line consists of all numbers from 500 up to 599. The last two digits indicate the train within the line. The odd numbers are used for one direction, the even numbers for the opposite direction. Usually even trains go towards Amsterdam. Low numbers are used for early trains, high numbers for later trains: two trains which are one hour apart have train numbers which differ by 4. Taken the even/odd principle into account this means that a frequency of two trains per hour fit into this rule, which also nicely spans one day ($24 \cdot 4 = 96$, just under 100).

An example can be found in Table 2.1, where the trains from Haarlem to Maastricht (leaving Haarlem 11 minutes past the hour) are numbered 817, 821, 825, up to 889. The trains from Maastricht to Haarlem have the numbers 818, 822, 826, up to 886. If this line would run twice per hour, one would also have the numbers 819, 823, and so on, and 816, 820, and so on. However, exactly 30 minutes apart from the 800-line, the 900-line runs from Haarlem to the south, leaving Haarlem 41 minutes past the hour. The majority of the line (from Haarlem to Sittard) is the same, but the 900-line runs to Heerlen instead of Maastricht. This leads to the train numbers 919, 923, 927, and so on, and 916, 920, 924, and so on. Below, the lines are given for the odd direction.

Some peak-hour lines which operate more than twice per hour per direction (marked ^{a}) deviate from the 4-points per hour rule. The international lines (marked ^{b}) are also numbered differently. The official abbreviations used for the stations can be found in Appendix A.3.

line-number	type[1]	frequency per hour[2] p	op	route, main stations[3]	remark
100^b	ICE	1*d*		Asd - Ut - Ah - Koln - Frank - Basel SBB	*only 105, 106*
120^b	ICE	6*d*		Asd - Ut - Ah - Koln - Frank	*only 120-129, 220-223*
140^b	IC	4*d*		Asd - Amf - Hgl - Hann - Berlin Ostbahnhof	*only 140-147*
300^b	CNL	1*d*		Asd - Ut - Ah - Köln - Basel SBB - Zürich HB	*only 318-319*
400^b	IC	1*d*		Rtd - Hld	*only 452, 453, 455; connecting to boat service to Harwich*
500	IC	1	1	Gvc - Ut - Amf - Zl - Gn	*see also 10500, 12500, 20500*
600	IC	1	1	Asd - Shl - Gv - Rtd - Rsd - Berch - Brusz	
700	IC	1	1	Shl - Amf - Zl - Gn	*see also 10700, 20700*
800	IC	1	1	Hlm - Asd - Ut - Ht - Ehv - Std - Mt	
900	IC	1	1	Hlm - Asd - Ut - Ht - Ehv - Std - Hrl	
1400	IC	-	*	Ut - Asd - Shl - Gvc - Rtd	*only operates 1am-6am*

1600	IC	1	1	Shl - Amf - Apd - Dv - Aml - Hgl - Es	*see also 21600*
1700	IC	1	1	Gvc - Ut - Amf - Apd - Dv - Aml - Hgl - Es	*see also 17000, 21700*
1900	IC	1	1	Gvc - Gv - Rtd - Ddr - Bd - Tb - Ehv - Vl	
2000	IC	2	2	Gvc - Gd - Ut - Ah [- Nm]	*see also 2800, 22000*
2100	IC	1	1	Asd - Shl - Ledn - Gv - Rtd - Ddr - Rsd - Vs	
2200	IR	2(2)	2(1)	Asd - Hlm - Ledn - Gv - Rtd - Ddr (- Bd)	
2400	IC	1	1	Asd - Shl - Ledn - Gv - Rtd - Ddr - Rsd	
2500	IC	1	1	Gvc - Gv - Ddr - Bd - Tb - Ehv	
2600	IR	2	2	Asd - Shl - Ledn - Gvc	
2700[b]	R	1	1	Rsd - Atw	
2800	IR	2	-	[Rtd - Gd - Ut]	*instead of 22000; see also 2000, 22000*
3000	IC	2	2	Hdr - Amr - Asd - Ut - Ah - Nm	
3200	IR	2	-	[Hnk - Hn - Asd]	*morning peak-hours only Hnk-Asd; evening peak-hours only Asd-Hnk*
3300	R	1	1	Hnk - Hn - Shl - Hfd	
3400	IR	2	2	Hn - Amr - Hlm - Ledn - Gvc	
3500	IC	2	2	[Asd -] Ut - Ht - Ehv	*exceptional services during peak-hours on Ehv-Rm-Std-Mt; see also 23500*
3600	IR	2	2	Zl - Dv - Zp - Ah - Nm - Ht - Tb - Bd - Rsd	
3800	IR	1	1	Zl - Emn	
3900	IR	2	2	Hfd - Shl - Asd - Lls	
4000	R	1	1	Asd - Gd - Rtd	
4100	R	2(1)	2(1)	(Hlds -) Hld - Rtd	
4200	R	2	2	Msw - Rtd	
4300	R	2	2	Hfd - Dvd - Lls	
4400	R	2	2	Nm - Ht	
4500	R	2	2	Ekz - Hn - Asd	
4600	R	2	2	Asd - Almb	
4700	R	2	2	Utg - Asd	
4800	S	2	2	Utg - Hlm - Asd	
4900	R	2(2)	2(1)	(Almb -) Ndb - Ut	
5000	R	2(2)	2(1)	(Ledn - Gv -) Rtd - Ddr	*see also 15000*
5100	R	2(2)	2(1)	Gvc - Gv - Rtd - Ddr (- Rsd)	
5200	R	2	2	Bd - Tb - Ehv	
5300[b]	R	1	1	Mt - Luik	
5500	R	2	2	Ut - Brn	
5600	IR	2	2	Ut - Amf - Zl	
5700	R	2	2	Ledn - Shl - Dvd - Hvs - Ut	
5800	IR	2	2	[Sgn -] Amr - Asd - Amf - Amfs	
5900	R	2	2	Ut - Rhn	
6000	R	2	2	Ut - Gdm - Tl	*see also 16000*
6100	R	1	1	Tl - Ah	
6200	IR	1	1	Nm - Vl - Rm	
6300	R	2	2	Zvt - Hlm - Ledn - Gvc	
6400	R	2	2	Ehv - Wt	
6500	R	1	1	Ehv - Vl	
6600	R	1	1	Mtr - Mt - Hrl - Krd	
6700	IR	1	1	Mtr - Mt - Hrl	

6800	R	2	2	Rm - Std - Mt - Mtr	
6900	R	2[1]	1	Std - Hrl [-Krd]	
7000	R	2	1	Dv - Aml	
7100	R	2(1)	2(2)	Ddr - Gr (-Gdm)	
7300	R	2	2	Amr - Asd - Ut	
7400	R	2	1	Ut - Amf - Ed	
7500	R	2	1	[Ut -] Ed - Ah	
7600	R	2	2	Nm - Ah - Zp	
8000	R	1	1	Zl - Emn	
8100	tram	2	2	Htn - Htnc	
8400[b]	R	7d		Nsch - Leer	
8500	R	2	2	Zl - Kpn	
8800	R	2	2	Ledn - Apn - Ut	
8900[b]	R	1	1	Hrl - Ahbf - Stola	
9000[b]	R	1	1	Vl - Dussel - Hamm	
9100	R	1	1	Zl - Gn	
9300[b]	Th	5*d*		Asd - Shl - Gv - Rtd - Berch - Brusz - Paris	
9500	LR	2	2	Apn - Gd	
9600	R	2	2	Ht - Ehv - Dn	
9700	R	2	-	Rtd - Gd - Gdg	
9800	R	2	2	Gvc - Gd - Gdg	
9900[b]	Th	1*w*		Asd - Shl - Gv - Rtd - Berch - Brusz - Marne-la-Vallée-Chessy	*only 9947, 9956*
9900[b]	Th	1*w*		Asd - Shl - Gv - Rtd - Berch - Brusz - Bourg-St-Maurice	*only 9920, 9987; only during winter*
9900[b]	Th	1*w*		Asd - Shl - Gv - Rtd - Berch - Brusz - Marseille-St-Charles	*only 9928, 9955, only during summer*
10500	IC	1	1	Zl - Lw	*decoupled from 500 (Zl); see also 500, 12500, 20500*
10700	IC	1	1	Zl - Lw	*decoupled from 700 (Zl); see also 700, 20700*
12500	IC	*	-	[Rtd - Ut - Amf - Zl - Lw]	*exceptional services during peak-hours, instead of 20500 and 10500; see also 500, 12500, 20500*
13300	R	2	2	Gvc - Ztmd - Gvc	*Zoetermeer-cityline, clockwise; changes train number –from odd to even– at Ztmd*
13400[a]	R	4	-	[Gvc - Ztmd - Gvc]	*Zoetermeer-cityline, clockwise; changes train number –from odd to even– at Ztmd*
13500	IR	2	-	[Gvc - Rth]	
13600	R	2	2	Gvc - Rth	
13700	R	2	2	Gvc - Zsh - Gvc	*Zoetermeer-cityline, counter-clockwise; changes train number –from odd to even– at Zsh*
13800[a]	R	4	-	[Gvc - Zsh - Gvc]	*Zoetermeer-cityline, counter-clockwise; changes train number –from odd to even– at Zsh*
14000	R	1	1	Ut - Gd - Rtd	
14100[a]	R	4	-	Vdg - Rtd	
15400	R	-	2	Zvt - Hlm - Asd	*only during summer months: Zvt-Asd only 14.30-19.30; Asd-Zvt only 9.30-15.00*
15000	R	2	2	Ledn - Gv	*runs only when 5000 does not; see also 5000*
16000	IC	2	2	Gdm - Ht	*decoupled from 6000 (Gdm); see also 6000*
16300	R	1	1	Br - Vl - Rm	

16500	R	1	1	Nm - Bmr	
16600	R	1	1	Nm - Vry	
17000	IC	*	-	[Rtd - Ut - Amf - Apd - Dv]	*exceptional services during peak-hours, instead of 21700; see also 17000, 21700*
17800	R	2	2	Apd - Zp	
18100	tram	2	2	Htn - Htnc	
18800	IR	2	-	[Ledn - Ut]	*morning peak-hours only Ledn-Ut; evening peak-hours only Ut-Ledn*
19500	R	2	-	[Ledn - Apn - Gd]	*morning peak-hours only Gd-Ledn; evening peak-hours only Ledn-Gd*
20500	IC	1	1	Rtd - Ut	*coupled with 500 (Ut); see also 500, 10500, 12500*
20700	IC	1	1	Asd - Amf	*coupled with 700 (Amf); see also 700, 10700*
21600	IC	1	1	Asd - Amf	*coupled with 1600 (Amf); see also 1600*
21700	IC	1	1	Rtd - Ut	*coupled with 1700 (Zl); see also 1700, 17000*
22000[4]	IC	2	2	Rtd - Gd	*coupled with 2000 (Gd); not when 2800 runs;see also 2000, 2800*
23500	IC	*	-	(Asd - Ut -) Ht - Tb - Bd - Rsd	*only 23522 Rsd - Asd instead of 3522; and 23561 Ht - Rsd (decoupled from 3561 (Ht)); see also 3500*
29000[b]	R	1	1	Es - G - Munst	*29004-29040*
29050[b]	R	1	1	Es - G - Dortmund Hbf	*29054-29084*
30000	R	2(1)	2(1)	Lw - Sk (- Stv)	
30100	R	2	2	Lw - Hlgh	
30200	R	2	2	Lw - Gn	
30300	IR	1	-	[Lw - Gn]	
30400	R	2(2)	2(1)	Gn - Ws (- Nsch)	
30500	R	2	1	Gn - Rd	
30600	R	2	2	Gn - Dz	
30700	R	2	-	[Ah - Dtc]	
30800	R	2	2	Zp - Ww	
30900	R	2	2	Ah - Dtc - Ww	
31000	R	2	1	Aml - Mrb	
31200	R	2	2	Zp - Hgl - Odz	

[1] CNL = CityNightLine; IC = intercity train; ICE = InterCity Express; IR = interregional train; LR = Light Rail; R = regional train; S = Sprinter; Th = Thalys.
[2] *p* indicates frequency during peak-hours; *op* is for off-peak-hours; numbers within parentheses (...) indicate frequency of the route-part within parentheses; numbers with *d* or *w* indicate the number of trains per day or week instead of the number of trains per hour.
[3] routes within parentheses (...) run less; routes within brackets [...] only during peak hours. Some lines operate less frequent, shorter routes, or not at all at night or during the weekend.
All lines are operated by NS, except for: 100 and 300 in cooperation with DB (German Railways) and SBB (Swiss Railways), 120 and 140 in cooperation with DB; 600 and 5300 in cooperation with NMBS (Belgian Railways); 2700 by NMBS; 8100, 9500 and 18100 in cooperation with HTM (The Hague bus and tram operator); 9300 and 9900 in cooperation with NMBS and SNCF (French Railways); 30000, 30100, 30200, 30300, 30400, 30500 and 30600 by NoordNed; 30700, 30800, 30900 and 31200 by Syntus; 31000 by Syntus by order of Connexxion; 8400, 8900, 9000, 29000 and 29050 by DB.

A.5 Punctuality Measurement

The delays of the railway traffic are measured at thousands of locations in the Netherlands. However, official delay figures are expressed in 3-minute arrival punctuality at the 34 large stations. These are represented in Figure A.1. The bold lines indicate the *core-network*, for which the official punctuality is measured. Under the new performance contract, signed by the national government and NS in April 2005, four local lines are added to the core-network: Amsterdam Central-Hoorn-Enkhuizen, Amsterdam Central-Almere Central-Lelystad Centrum, Utrecht-Baarn, and Utrecht-Rhenen. Hoorn is the 35$^{\text{th}}$ measuring station.

A.6 Original Dutch Names

For the readability of the text, English names are used for Dutch institutions and projects. The original Dutch names are given in this appendix.

Commission Brokx Commissie Brokx

Commission Wijffels Commissie Wijffels

High Speed Line-South (HSL-South) Hogesnelheidslijn-Zuid (HSL-Zuid)

Ministry of Transport Ministerie van Verkeer en Waterstaat (V&W)

Rail Infrastructure Management RailInfrabeheer

Rail Traffic Control RailVerkeersleiding

Second scheme for the structure of traffic and transport Tweede Structuurschema Verkeer en Vervoer

Utilize and Build Benutten en Bouwen

A.7 Software

For the readability of the text, English names are used for Dutch institutions and projects. The original Dutch names are given in this appendix.

CPLEX (version 9.0) Optimizer for mathematical programs

DONS (Designer of Network Schedules, version 8.5) Semi-automatic construction of cyclic railway timetables

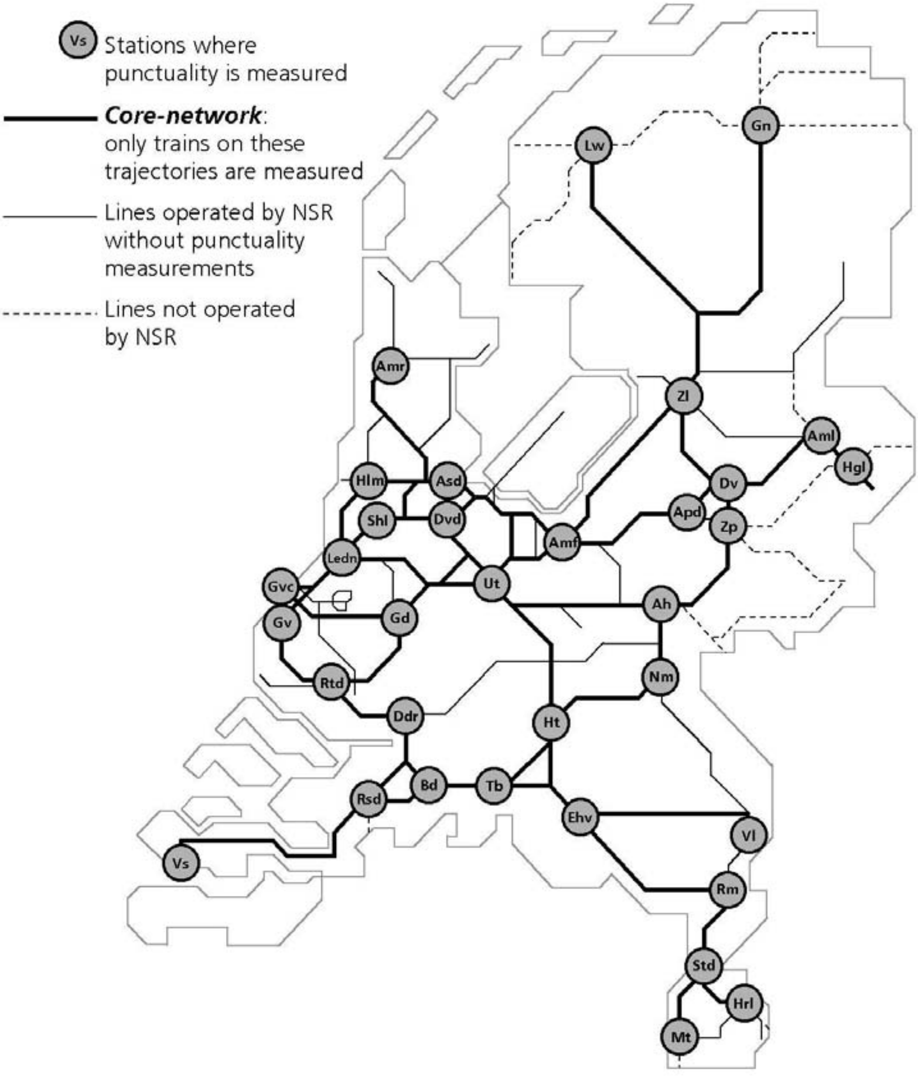


Figure A.1: The official punctuality measurements take place at the indicated stations.

FRISO (Flexibele Rail Infra Simulatie Omgeving, or, in English, Flexible Rail Infrastructure Simulation Environment) simulation of railway systems on a detailed level

OPL Studio (Optimization Programming Language Studio, version 3.7) Modeling environment based on the OPL programming language

PETER (Performance Evaluation of Timed Events in Railways, version 2.2.6) Timetable evaluation based on max-plus algebra

SIMONE (SImulation MOdel of NEtworks, production version 6.1.5) Simulation of railway systems on a network wide basis

Appendix B

Case Haarlem–Maastricht/Heerlen

The Timetable of the Reference Case

This appendix describes the cyclic timetable for the Haarlem–Maastricht/Heerlen case. This timetable is applicable to both the original and the optimized situation in Section 4.4. This is because the changes were not made in the planned running times, but, implicitly, in the minimal running times. Table B.1 shows the southbound trains, and Table B.2 shows the northbound trains. The southbound timetable is also used for the stochastic optimization model in Section 6.5.

The departure and arrival times are only given for the stations where the particular train line dwells. Departures are indicated with a 'd', arrivals with an 'a'. Many stops are not indicated by a 'd' or an 'a'. Here the arrival and departure times are equal: these dwells are scheduled to be zero. This is exactly how the simulation interprets these stops. The DONS-system, with which the timetable was created, includes short stops in the running time preceding this stop.

Note that train lines 6000 (Ut-Ht) and 9600 (Ht-Ehv) are considered to be one line (Ut-Ehv) in Sections 4.4 and 6.5. This is in fact the real life situation, because the rolling stock arriving as 6000 in 's-Hertogenbosch (Ht) continues as the 9600 from 's-Hertogenbosch. The same is true for the opposite direction.

Train lines 2200, 3000, 3500, 4800, 6000, 6400, 6800, 6900, 7300, 9600 and 19600 run twice per hour. The second train of these lines runs exactly 30 minutes apart from the one presented.

The horizontal lines in the tables represent the delay measuring locations. Furthermore, in Figure 4.12 it can be seen that Sittard–Maastricht (Std-Lut-Bk-Bde-Mt) and Sittard–Heerlen (Std-Gln-Sbk-Sn-Nh-Hb-Hrl) are two disjoint trajectories.

event		ICE 100	IC 800	IC 900	IR 2200	IC 3000	IC 3500	R 4000	R 4800	R 6000	R 6400	R 6800	R 6900	R 7300	R 9600	R 19600
Hlm	d		54	24	20				10							
Hlms									14							
Ass					30				21							
Asd	a		08	38	35				26							
Asd	d	16	10	40		28	20	02						23		
Asdm								08						29		
Asa			18	48		36		10						31		
Dvd	a	25	21	51		39	30	14						35		
Dvd	d	26	22	52		40	31	14						35		
Asb							34	16						37		
Ac	a							21						42		
Ac	d							21						46		
Bkl								29						54		
Mas														58		
Ut	a	43	39	09		57	54							05		
Ut	d		41	11			56			02						17
Utl										07						
Htn										11						24
Cl										18						31
Gdm	a									24						37
Gdm	d									34						
Zbm										40						
Ht	a		09	39			24			50						
Ht	d		11	41			25								27	
Vg															31	
Btl															37	
Bet															44	
Ehb															49	
Ehv	a		29	59			43								53	
Ehv	d		31	01							07					
Gp											14					
Hze											18					
Wt			48	18							29					
Rm	a		02	32												
Rm	d		03	33								07				
Ec												16				
Srn												20				
Std	a		18	48								26				
Std	d		19	49								27	23			
Lut												31				
Bk												35				
Bde												41				
Mt	a		34									47				
Gln				55									28			
Sbk													32			
Sn													34			
Nh													38			
Hb													41			
Hrl	a			07									45			

Table B.1: The departure and arrival times of the lines on the Haarlem–Maastricht/Heerlen corridor, southbound.

event		ICE 100	IC 800	IC 900	IR 2200	IC 3000	IC 3500	R 4000	R 4800	R 6000	R 6400	R 6800	R 6900	R 7300	R 9600	R 19600
Hrl	d			28									49			
Hb													54			
Nh													57			
Sn													01			
Sbk													03			
Gln				40									07			
Mt	d		00									18				
Bde												24				
Bk												30				
Lut												34				
Std	a		16	46								38	12			
Std	d		17	47								39				
Srn												45				
Ec												49				
Rm	a		32	02								58				
Rm	d		33	03												
Wt			47	17							06					
Hze											17					
Gp											21					
Ehv	a		03	33							27					
Ehv	d		05	35			22								12	
Ehb															15	
Bet															21	
Btl															29	
Vg															35	
Ht	a		24	54			41								39	
Ht	d		26	56			42			16						
Zbm										26						
Gdm	a									32						
Gdm	d									42						29
Cl										48						35
Htn										55						42
Utl										59						
Ut	a		54	24			10			04						49
Ut	d	22	56	26		09	12							03		
Mas														10		
Bkl								08						13		
Ac	a							16						21		
Ac	d							16						25		
Asb							34	20						29		
Dvd	a	39	14	44		27	36	23						32		
Dvd	d	40	15	45		28	37	23						32		
Asa			19	49		32		27						36		
Asdm								30						39		
Asd	a	50	27	57		40	47	35						44		
Asd	d		29	59	02				11							
Ass					08				16							
Hlms									24							
Hlm	a		43	13	17				27							

Table B.2: The departure and arrival times of the lines on the Maastricht/Heerlen–Haarlem corridor, northbound.

Fixed Departure and Arrival Times

Some of the departure and arrival times of the timetable above are fixed for the optimization cases in Section 6.5. This is to ensure that the optimized timetable fits within the national timetable. The lines and locations that are concerned are tabled for both directions in Table B.3.

	southbound				northbound			
line	event		time(s)		event		time(s)	
100	Ut	a	43		Ut	d	22	
2200	Hlm	d	20	50	Hlm	a	17	47
3000	Asd	d	28	58	Ut	d	09	39
3000	Ut	a	57	27	Asd	a	40	10
4000	Bkl	a	29		Bkl	d	08	
4800	Hlm	d	10	40	Hlm	a	27	57
7300	Asd	d	23	53	Asd	a	14	44
9600	Ehv	a	23	53	Ehv	d	12	42

Table B.3: Fixed departure and arrival times for the optimization of the corridor, southbound.

Bibliography

Abbink, E.J.W., Fischetti, M., Kroon, L.G., Timmer, G., Vromans, M.J.C.M. (2004). *Reinventing crew scheduling at Netherlands Railways (URL = http://hdl.handle.net/1765/1427)*. Technical report, ERS-2004-046-LIS, Erasmus University Rotterdam, Rotterdam, the Netherlands.

Al-Ibrahim, A., Oldenziel, M.J. (2004). *Capacity management. Simulation study Dordrecht-Roosendaal-Breda*. Technical report, ProRail, Utrecht, the Netherlands.

Apostol, T.M. (1967). *Calculus*, volume I. John Wiley and Sons, Chichester, United Kingdom, 2nd edition.

Assad, A.A. (1980). *Models for rail transportation*. Transportation Research Part A, 14:205–220.

Baccelli, F., Cohen, G., Olsder, G.J., Quadrat, J.-P. (1992). *Synchronization and Linearity*. John Wiley and Sons, Chichester, United Kingdom.

Bergmark, R. (1996). *Railroad capacity and traffic analysis using SIMON*. In Allan, J., Brebbia, C.A., Hill, R.J., Sciutto, G. (editors), *Computers in Railways V*, pages 183–191. WIT Press, Southampton, United Kingdom.

Billionnet, A. (2003). *Using integer programming to solve the train-platforming problem*. Transportation Science, 37(2):213–222.

Birge, J.R., Louveaux, F. (1997). *Introduction to stochastic programming*. Springer, New York, New York, United States.

Bouwman, M.R., Joustra, P.E., Teunisse, M.J.G. (2004). *Conceptstudie ontwikkeling stabiliteitsmodel voor railnetwerken (Concept study development stability model for rail networks)*. Incontrol Enterprise Dynamics, Maarssen, the Netherlands. In Dutch.

Braker, J.G. (1993). *Algorithms and applications in timed discrete event systems*. Ph.D. thesis, Technical University Delft, Delft, the Netherlands.

Braker, J.G., Olsder, G.J. (1993). *The power algorithm in max algebra*. Linear Algebra and its applications, 182:67–89.

Bruinsma, F.R., Rietveld, P., van Vuuren, D.J. (1998). *Unreliability in public transport chains*. Technical report, Tinbergen Institute Discussion Paper, Amsterdam, the Netherlands.

Caprara, A., Fischetti, M., Toth, P. (2002). *Modeling and solving the train timetabling problem*. Operations Research, 50(5):851–861.

Carey, M. (1998). *Optimizing scheduled times, allowing for behavioural response*. Transportation Research Part B, 32(5):329–342.

——— (1999). *Ex ante heuristic measures of schedule reliability*. Transportation Research Part B, 33(7):473–494.

Carey, M., Carville, S. (2000). *Testing schedule performance and reliability for train stations*. Journal of the Operational Research Society, 51(6):666–682.

——— (2003). *Scheduling and platforming trains at busy complex stations*. Transportation Research Part A, 37(3):195–224.

Carey, M., Kwieciński, A. (1995). *Properties of expected costs and performance measures in stochastic models of scheduled transport*. European Journal of Operational Research, 83:182–199.

Cochet-Terrasson, J., Cohen, G., Gaubert, S., Mc Gettrick, M., Quadrat, J.-P. (1998). *Numerical computation of spectral elements in max-plus algebra*. IFAC Conference on System Structure and Control, Nantes, France.

Cochran, W.G. (1977). *Sampling Techniques*. John Wiley and Sons, New York, New York, United States, 3rd edition.

Commission Brokx (1995). *Marktwerking in het openbaar vervoer (Privatization in public transport)*. In Dutch.

Cordeau, J.-F., Toth, P., Vigo, D. (1998). *A survey of optimization models for train routing and scheduling*. Transportation Science, 32(4):380–404.

De Klerck-Salm, S. (2004). *PRRING. Case study Den Bosch*. Technical report, ProRail, Utrecht, the Netherlands. In Dutch.

De Kort, A.F. (2000). *Advanced railway planning using (max,+) algebra*. In Allan, J., Hill, R.J., Brebbia, C.A., Sciutto, G., Sone, S. (editors), *Computers in Railways VII*, pages 257–266. WIT Press, Southampton, United Kingdom.

De Kort, A.F., Heidergott, B., Ayhan, H. (2003). *A probablistic (max,+) approach for determining railway infrastructure capacity*. European Journal of Operational Research, 148:644–661.

Dekker, R., Vromans, M.J.C.M. (2001). *Punctualiteit in het reizigersvervoer per trein (Punctuality in passenger transport by train)*, pages 176–189. Stichting Toekomst der Techniek, The Hague, the Netherlands. In Dutch.

Demitz, J., Hübschen, C., Albrecht, C. (2004). *Timetable stability – using simulation to ensure quality in a regular interval timetable*. In Allan, J., Brebbia, C.A., Hill, R.J., Sciutto, G., Sone, S. (editors), *Computers in Railways IX*, pages 549–562. WIT Press, Southampton, United Kingdom.

EC (European Community) (1991). *Directive 91/440/EEC*. Official Journal of the European Communities.

Feller, W. (1977). *An Introduction to Probability Theory and Its Applications, Volume 1*. John Wiley and Sons, New York, New York, United States, 3rd edition.

Fischer, H., Bolemant, M., Margaritov, A., Friedrich, G.-R. (1995). *Das Projekt SABINE - Simulation und Auswertung des Betriebsablaufs im Netz (The SABINE project - Simulation and handling of the operations in the network)*. Simulation und Simulatoren für den Schienenverkehr, pages 451–462. In German.

Ghosh, J.B. (1984). *Asynchronous simulation of some discrete time models*. In Sheppard, S., Pooch, U., Pegden, D. (editors), *Proceedings of the 16th conference on Winter simulation*, pages 466–469. Institute of Electrical and Electronic Engineers Press, Piscataway, New Jersey, United States.

Gielissen, A.N.J. (1983). *Interpretatie van rijtijden (Interpretation of running times)*. Internal document, NS, Utrecht, the Netherlands.

Google (2005). *Definitions of reliability on the web*. Internet: http://www.google.nl/search?hl=nl&lr=&oi=defmore&q=define:Reliability. Visited: January 2005.

Goossens, J.H.M. (2004). *Models and algorithms for railway line planning models (URL = http://www.t75.nl/jgoossens/files/thesis-hyperref.zip)*. Ph.D. thesis, University of Maastricht, Maastricht, the Netherlands.

Goverde, R.M.P. (1998). *Optimal scheduling of connections in railway systems*. Presented at the 8th World Congress on Transport Research (WCTR), Antwerp, Belgium.

——— (1999). *Improving punctuality and transfer reliability by railway timetable optimization*. In *Proceedings TRAIL 5th Annual Congress*, volume 2. Delft University Press.

——— (2002). *Stability analysis of periodic railway timetables (A max-plus algebra approach)*. In *Proceedings TRAIL 7th Annual Congress*. Delft University Press, Delft, the Netherlands.

——— (2003). *Bruikbaarheid en kwaliteit van realisatiegegevens treinactiviteiten (Usability and quality of train activity realization data)*. Technical report, ProRail, Utrecht, the Netherlands. In Dutch.

Goverde, R.M.P., Hooghiemstra, G., Lopuhaä, H.P. (2001). *Statistical analysis of train traffic; the Eindhoven case*. Technical report, TRAIL, Technical University Delft, Delft, the Netherlands.

Goverde, R.M.P., Odijk, M.A. (2002). *Performance evaluation of network timetables using PETER*. In Allan, J., Hill, R.J., Brebbia, C.A., Sciutto, G., Sone, S. (editors), *Computers in Railways VIII*, pages 731–740. WIT Press, Southampton, United Kingdom.

Goverde, R.M.P., Soto y Koelemeijer, G. (2000). *Performance evaluation of periodic railway timetables: theory and algorithms*. Technical report, TRAIL, Technical University Delft, Delft, the Netherlands.

Gröger, T. (2004). *Timetable simulation by sophisticated conflict resolution between train paths*. In Allan, J., Brebbia, C.A., Hill, R.J., Sciutto, G., Sone, S. (editors), *Computers in Railways IX*, pages 573–582. WIT Press, Southampton, United Kingdom.

Haldeman, L. (2003). *Automatische Analyse von IST-Fahrplnen (Automatic analyses of operational timetables)*. Master's thesis, ETH Zürich, Zürich, Switzerland. In German.

Hansen, I.A. (2000). *Station capacity and stability of train operations*. In Allan, J., Hill, R.J., Brebbia, C.A., Sciutto, G., Sone, S. (editors), *Computers in Railways VII*, pages 853–860. WIT Press, Southampton, United Kingdom.

Harrod, S. (2003). *Timetable buffer magnitude and allocation; an analysis using simulation*. Presented at 2003 INFORMS National Meeting, Atlanta, Georgia, United States.

Hendriks, P. (2004). *Effect van planningnormen en lijnvoeringen op punctualiteit (Effect of planning norms and line plans on punctuality)*. Internal document, ProRail, Utrecht, the Netherlands. In Dutch.

Higgins, A., Kozan, E. (1998). *Modeling train delays in urban networks*. Transportation Science, 32(4):346–357.

Higgins, A., Kozan, E., Ferreira, L. (1995). *Modeling delay risks associated with train schedules*. Transportation Planning and Technology, 19:89–108.

Hooghiemstra, J.S. (1994). *An automatic timetable generator for the Dutch railway network*. In Allan, J., Brebbia, C.A., Hill, R.J., Sciutto, G. (editors), *Computers in Railways IV*, volume 2, pages 109–116. Computational Mechanics, Southampton, United Kingdom.

——— (1996). *Design of regular interval timetables for strategic and tactical railway planning*. In Allan, J., Brebbia, C.A., Hill, R.J., Sciutto, G. (editors), *Computers in Railways V*, volume 1, pages 393–402. WIT Press, Southampton, United Kingdom.

Hooghiemstra, J.S., Kroon, L.G., Odijk, M.A., Salomon, M., Zwaneveld, P.J. (1999). *Decision Support Systems support the search for win-win solutions in railway network design*. Interfaces, 29(2):15–32.

Huisman, D., Kroon, L.G., Lentink, R.M., Vromans, M.J.C.M. (2005). *Operations research in passenger railway transportation*. Technical report, EI2005-16, Econometric Institute, Erasmus University Rotterdam, Rotterdam, the Netherlands.

Huisman, T. (2005). *Queueing models for railway systems*. Ph.D. thesis, University of Twente, Enschede, the Netherlands.

Huisman, T., Boucherie, R.J. (2001). *Running times on railway sections with heterogeneous train traffic*. Transportation Research Part B, 35:271–292.

Huisman, T., Boucherie, R.J., van Dijk, N.M. (2002). *A solvable queueing network model for railway networks and its validation and applications for the Netherlands*. European Journal of Operational Research, 142:30–51.

Hürlimann, D. (2001). *Objektorientierte Modellierung von Infrastrukturelementen und Betriebsvorgängen im Eisenbahnwesen (Object oriented modeling of infrastructure elements and business processes in railways)*. Ph.D. thesis, ETH Zürich, Zürich, Switzerland. In German.

IEEE (Institute of Electrical and Electronics Engineers) (1990). *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*. New York, New York, United States.

Kaas, A.H. (2000). *Punctuality model for railways*. In Allan, J., Hill, R.J., Brebbia, C.A., Sciutto, G., Sone, S. (editors), *Computers in Railways VII*, pages 809–816. WIT Press, Southampton, United Kingdom.

Karp, R.M. (1978). *A characterization of the minimum cycle mean in a digraph*. Discrete Mathematics, 23:309–311.

Klein Haneveld, W.K., Van der Vlerk, M. (2004). *Stochastic programming*. Lecture notes. Rijksuniversiteit Groningen, Groningen, the Netherlands.

Knuth, D.E. (1997). *The Art of Computer Programming, 3rd edition, Vol. 2*. Addison Wesley Longman, Boston, Massachusetts, United States.

König, H. (2001). *VirtuOS, Simulieren von Bahnbetrieb (VirtuOS, simulation of railways)*. Betrieb und Verkehr, 50(1–2):44–47. In German.

Kroon, L.G. (2001). *Opsporen van sneller naar beter; modelling through (trekking from faster to better; modelling through)*. Inaugural address. Erasmus University Rotterdam, Rotterdam, the Netherlands. In Dutch.

Kroon, L.G., Peeters, L.W.P. (2003). *A variable trip time model for cyclic railway timetabling*. Transportation Science, 37:198–212.

Law, A.M., Kelton, W.D. (2000). *Simulation modeling and analysis*. McGraw-Hill, Chichester, United Kingdom, 3rd edition.

Liebchen, C. (2003). *Symmetry for periodic railway timetables*. Electronic notes in theoretical computer science, 92(1).

Linderoth, J.T., Shapiro, A., Wright, S.J. (2002). *The emprirical behavior of sampling methods for stochastic programming*. Optimization Technical Report 02-01, University of Wisconsin-Madison, (2002), Madison, Wisconsin, United States.

Loeve, L., Tressel, J., Smoltsak, I. (2005). *FRISO conceptueel model (FRISO conceptual model)*. Incontrol Enterprise Dynamics, Maarssen, the Netherlands. In Dutch.

Middelkoop, D., Bouwman, M. (2000). *Train network simulator for support of network wide planning of infrastructure and timetables*. In Allan, J., Hill, R.J., Brebbia, C.A., Sciutto, G., Sone, S. (editors), *Computers in Railways VII*, pages 267–276. WIT Press, Southampton, United Kingdom.

——— (2001). *SIMONE: large scale train network simulations*. In Peters, B., Smith, J. (editors), *Proceedings of the 2001 Winter Simulation Conference*, pages 1042–1047. Institute of Electrical and Electronic Engineers Press, Piscataway, New Jersey, United States.

Nachtigall, K. (1996). *Periodic network optimization with different arc frequencies*. Discrete Applied Mathematics, 69:1–17.

——— (1998). *Periodic network optimization and fixed interval timetables*. Habilitationsschrift, Universität Braunschweig, Germany.

Nachtigall, K., Voget, S. (1997). *Minimizing waiting times in integrated fixed interval timetables by upgrading railway tracks*. European Journal of Operational Research, 103:610–627.

Nordbeck, H. (2003). *Het rijdend personeel laat van zich horen. Hoe het rondje om de kerk verdween. (The train crew repels. How the church circles disappeared)*. Utrecht, the Netherlands. In Dutch.

NS (Nederlandse Spoorwegen) (1988). *Rail 21: sporen naar een nieuwe eeuw (Rail 21: trekking to a new century)*. Utrecht, the Netherlands. In Dutch.

——— (1990). *Rail 21: een nadere onderbouwing (Rail 21: a further foundation)*. Utrecht, the Netherlands. In Dutch.

——— (1992). *Rail 21: het treindienst concept voor de volgende eeuw (Rail 21: timetable concepts for the next century)*. Utrecht, the Netherlands. In Dutch.

——— (1994). *Scenario studie Rail 21 (Scenario study Rail 21)*. Utrecht, the Netherlands. In Dutch.

——— (2000). *De Koppeling, nr.1619, 12-16-2001*. Utrecht, the Netherlands. In Dutch.

——— (2002). *Spoorboekje, 15 december 2002 t/m 13 december 2003 (Timetable 2003)*. Utrecht, the Netherlands. In Dutch.

——— (2004a). *Annual Report 2003*. Utrecht, the Netherlands.

——— (2004b). *Klanttevredenheids onderzoek (Passenger satisfaction inquiry); Internet: http://www.ns.nl/over+ns/over+ns.cgi?Over+NS/Cijfers/Kwaliteitscijfers*. In Dutch.

——— (2004c). *NS Home; Internet: http://www.ns.nl/*. In Dutch.

——— (2004d). *Wachttijd ReizigersTreinen (WRT), dienstregeling 2005*. Utrecht, the Netherlands.

NS (Nederlandse Spoorwegen), ProRail, Railion, V&W (Ministry of Transport) (2003). *Benutten en Bouwen, het plan van de spoorsector (Utilize and Build, the plan of the railway sector)*. Utrecht, the Netherlands. In Dutch.

Olde Hartman, F. (2003). *Modellen uit de luchtvaart op het spoor! (Aviation models in railways)*. Master's thesis, Universiteit Twente, Enschede, the Netherlands. In Dutch.

Olsson, N., Haugland, H. (2004). *Influencing factors on train punctuality–results from some Norwegian studies*. Transport Policy, 11:387–397.

Pachl, J. (2002). *Railway Operation and Control*. VTD Rail Publishing, Mountlake Terrace, Washington, United States.

Palm, H. (2004). *Betrouwbaarheid op het spoor (Reliability on the track)*. Tijdschrift Vervoerswetenschappen, 40(4):2–6. In Dutch.

Peeters, L.W.P. (2003). *Cyclic railway timetable optimization (URL = https://ep.eur.nl/handle/1765/429)*. Ph.D. thesis, Rotterdam School of Management, ERIM Ph.D. Series Research in Management, Rotterdam, the Netherlands.

Peeters, L.W.P., Kroon, L.G. (2001). *A cycle based optimization model for the cyclic railway timetabling problem*. In *Computer aided scheduling of public transport*, pages 275–296. Springer, Berlin, Germany.

Poort, J.P. (2002). *Grenzen aan benutting (Limits on utilization)*. NYFER, Breukelen, the Netherlands. In Dutch.

Rietveld, P., Bruinsma, F., van Vuuren, D., Rooijers, A.J., Peeters, P.M. (1998). *Hoe laat denk je thuis te zijn? Een onderzoek naar de betrouwbaarheid van vervoersystemen en de invloed daarvan op het verplaatsingsgedrag (When do you expect to be home? Researching the reliability of transport systems and its influence on travel habits)*. Onderzoeksrapport, Vrije Universiteit, vakgroep Ruimtelijke economie, Centrum voor Omgevings- en Verkeerspsychologie, Peeters Advies, Amsterdam, the Netherlands. In Dutch.

Rietveld, P., Bruinsma, F.R., van Vuuren, D.J. (2001). *Coping with unreliability in public transport chains: a case study for the Netherlands*. Transportation Research Part A, 35:539–559.

Rudolph, R. (2003). *Allowances and margins in railway scheduling*. In *Proceedings of World Congress on Railway Research 2003*, pages 230–238. Edinburgh, United Kingdom.

——— (2004). *Entwicklung von strategien zur optimierten anordnung und dimensionierung von zeitzuschlägen im eisenbahnbetrieb (Development of strategies to optimize the allocation and size of supplements in railway systems)*. Ph.D. thesis, Universität Hannover, Hannover, Germany. In German.

RUDOLPH, R., DEMITZ, J. (2003). *Simulation of large railway networks. Robustness test for the timetable 2003 in North Rhine Westphalia, Germany*. In *Proceedings of World Congress on Railway Research 2003*, pages 644–652. Edinburgh, United Kingdom.

SCHAAFSMA, A.A.M. (2001). *Dynamisch railverkeersmanagement (Dynamic rail traffic management)*. Ph.D. thesis, Technical University Delft, Delft, the Netherlands. In Dutch.

SCHRIJVER, A., STEENBEEK, A. (1994). *Dienstregelingontwikkeling voor Railned (Timetable construction for Railned)*. Technical report, Center for Mathematics and Computer Science, Amsterdam, the Netherlands. In Dutch.

SCHULZ VAN HAEGEN-MAAS GEESTERANUS, M.H. (2004). *Verhouding Rijksoverheid - NS (Relation national government - NS)*. Letter of the vice minister of traffic and waterworks to the house of representatives (0304tkkst18986-101). The Hague, the Netherlands. In Dutch.

SCHWANHÄUSSER, W. (1994). *The status of German railway operations management in research and practice*. Transportation Research Part A, 28(6):495–500.

SERAFINI, P., UKOVICH, W. (1989). *A mathematical model for Periodic Event Scheduling Problems*. SIAM Journal of Discrete Mathematics, 2(4):550–581.

SOTO Y KOELEMEIJER, G., IOUNOUSSOV, A.R., GOVERDE, R.M.P., VAN EGMOND, R.J. (2000). *PETER, a performance evaluator for railway timetables*. In ALLAN, J., HILL, R.J., BREBBIA, C.A., SCIUTTO, G., SONE, S. (editors), *Computers in Railways VII*, pages 405–414. WIT Press, Southampton, United Kingdom.

SUBIONO (2000). *On classes of min-max-plus systems and their applications*. Ph.D. thesis, Technical University Delft, Delft, the Netherlands.

SUBIONO, OLSDER, G.J., MC GETTRICK, M. (1998). *On large schale max-plus algebra model in railway systems*. Fourth International Workshop on Discrete Event Systems (WODES'98), Cagliari, Italy.

TIJMS, H.C. (2003). *A first course in stochastic models*. John Wiley and Sons, Chichester, United Kingdom.

UIC (INTERNATIONAL UNION OF RAILWAYS) (2000). *Timetable recovery margins to guarantee timekeeping - recovery margins*. Leaflet 451-1. Paris, France.

——— (2004). *Capacity*. Leaflet 406. Paris, France.

VAN EGMOND, R.J. (1999). *Railway capacity assesment, an algabraic approach*. Technical report, TRAIL, Technical University Delft, Delft, the Netherlands.

VAN NES, R. (1997). *Design of multimodal transport networks, a hierarchal approach*. Ph.D. thesis, Technical University Delft, Delft, the Netherlands.

VROMANS, M.J.C.M., KROON, L.G. (2004). *Stochastic optimization of railway timetables*. In *Proceedings TRAIL 8th Annual Congress*, pages 423–445. Delft University Press, Delft, the Netherlands.

VROMANS, M.J.C.M., KROON, L.G., DEKKER, R. (2006). *Reliability and heterogeneity of railway services*. European Journal of Operational Research.

V&W (MINISTRY OF TRANSPORT) (1988). *Tweede Structuurschema Verkeer en Vervoer (Second Structurescheme Traffic and Transport)*. The Hague, the Netherlands. In Dutch.

——— (2000). *Van A naar Beter, Nationaal Verkeers- en Vervoersplan (From A to Better, National Traffic and Transport Plan).* The Hague, the Netherlands. In Dutch.

WAHLBORG, M. (1996). *Simulation models: important aids for Banverket's planning process*. In *Computers in Railways V*, volume 1, pages 175–181. WIT Press, Southampton, United Kingdom.

WAKOB, H. (1985). *Ableitung eines generellen wartemodells zur ermittlung der planmäßigen wartezeiten im eisenbahnbetrieb unter besonderer berücksichtigung der aspekte leistungsfähigkeit und anlagebelastung (Derivation of a generic queueing model for the analysis of planned waiting times in railway systems).* Ph.D. thesis, Verkehrwissenschaftlichen Institut, RWTH Aachen, Aachen, Germany. In German, English summary.

WEIGAND, W. (1981). *Verspätungsübertragungen in fernverkehrsnetzen (Delay propagation in long distance traffic networks).* Eisenbahntechnische Rundschau, 30(12):915–920. In German.

WENDLER, E. (1995). *Influences of modern signalling systems on railway capacity*. In *Proceedings of the 2nd Erasmus-network conference on transportation and traffic engeneering*, pages 50–60. Kerkrade, the Netherlands.

WIJFFELS, H.H.F., IN HET VELD, R.J., DE SOET, J.F.A. (1992). *Sporen voor straks: advies over de toekomstige relatie tussen de overheid en Nederlandse Spoorwegen (Tracks for later: advice for the future relation between government and NS).* The Hague, the Netherlands. In Dutch.

WOJTKOWSKI, A. (2004). *Organisch plannen; Voorspelbaar betrouwbaar? (Orchanic planning; Predictably reliable?).* Technical report, ProRail, Utrecht, the Netherlands. In Dutch.

YBEMA, R. (2000). *Een verdeelde variant op de route IJssel-Brabant (A differentiated scenario on the IJssel-Brabant corridor).* Bachelor's thesis, Hogeschool Holland Diemen, Diemen, the Netherlands. In Dutch.

YUAN, J. (2001). *Statistical analyses of train delays at The Hague HS*. In HANSEN, I.A. (editor), *Train delays at stations and network stability (workshop).* TRAIL, Delft, the Netherlands.

YUAN, J., HANSEN, I.A. (2004). *Analyses of scheduled and real capacity utilization of a major Dutch railway station.* In ALLAN, J., BREBBIA, C.A., HILL, R.J., SCIUTTO, G., SONE, S. (editors), *Computers in Railways IX*, pages 595–602. WIT Press, Southampton, United Kingdom.

ZWANEVELD, P.J., KROON, L.G., AMBERGEN, H.W. (1996a). *A decision support system for routing trains through railway stations*. In ALLAN, J., BREBBIA, C.A., HILL, R.J., SCIUTTO, G. (editors), *Computers in Railways V*, volume 1, pages 217–226. WIT Press, Southampton, United Kingdom.

ZWANEVELD, P.J., KROON, L.G., ROMEIJN, H.E., M.SALOMON, DAUZÈRE-PÉRÈS, S., VAN HOESEL, C.P.M., AMBERGEN, H.W. (1996b). *Routing trains through railway stations: model formulation and algorithms*. Transportation Science, 30(3):181–194.

Summary

Reliability of Railway Systems

The utilization of the Dutch railway infrastructure has to increase to keep the urban areas accessible at acceptable costs. The railway network in the Netherlands is already one of the highest utilized networks in the world, but the Dutch railway industry is nonetheless aiming for an even higher utilization of the capacity. Railway traffic has to be much more reliable to achieve this goal.

Additionally, the on-time running of the trains itself is the most important goal of Netherlands Railways, it is an important characteristic to attract customers to railway transport. Moreover, the on-time running is also a central issue in contracts with the government and passenger organizations.

The main goal of this thesis is *"to develop rules and instruments for supporting the generation of more reliable timetables"*.

Chapter 1 provides a picture of the railway world in which this research has taken place. This is followed by the description of some concepts that are important to the reliability research. For example, *the 3-minute arrival punctuality* is the percentage of trains that arrives within three minutes from the planned arrival at a large station.

For the developments of these rules and models, it is first explained in Chapter 2 how the planning takes place at a railway company today. However, during the operation of the timetable there will be disturbances, due to for example malfunctioning infrastructure or rolling stock, or weather conditions. The resulting delays can propagate throughout the network because of the coherence of the railway system. Therefore, many of the dependencies within the railway system are described. Modeling a railway system, something which will repeatedly happen in later chapters, is the art of formalizing these dependencies.

Chapter 3 consists of the description of some important models for the construction and evaluation of timetables. First the *DONS*-software is introduced, which is used in the Dutch railway industry for the construction of cyclic timetables. Such a cyclic timetable, of 60 minutes, is repeated several times to create a timetable for a whole day.

Using *max-plus algebra*, some characteristics with respect to the reliability of cyclic timetables can be calculated.

The simulation software *SIMONE* is developed for the Dutch railway industry to judge timetables. Railway systems can be imitated in detail with simulation.

Railway and other public transport companies include running time supplements in their timetable to be able to absorb small delays. However, only limited research has looked into the best allocation of these supplements. Chapter 4 of this thesis describes this issue and concludes that the usual proportional allocation of the supplements in the Netherlands (everywhere 7% of the minimal running time) is not optimal. It is profitable to have less supplement on the first and last parts of a train line, and relatively somewhat more in the middle. This is the consequence of two intuitive causes. When a train uses supplement on the first part of its line, a delay reduction is achieved for all stations ahead. However, if the supplement is used on the last part of the line this is only beneficial for the delay reduction at the terminal station. On the other hand, too much supplement at the start will often lead to the situation that it is wasted, because the train did not incur any delay yet. The punctuality can be improved considerably by planning relatively much supplement halfway the train line.

Speed differences of differentiated train services lead, foremost on crowded trajectories, to small headways. The probability of delay propagation is much larger in case of small headways. The headways can be increased by *homogenizing* the railway traffic, i.e. reducing the speed differences. Consequently the delay propagation decreases. These dependencies are quantified by the two new measures SSHR and SAHR in Chapter 5. These measures do not only quantify the heterogeneity of the traffic, but they also provide an indication for the delay propagation. These results are supported by two extensive simulation studies.

Many of the dependencies within a railway system can be captured by the model presented in Chapter 6. The stochastic optimization model, a linear programming model that is deduced from so called recourse models, consists of two parts. The first part takes care of all the prerequisites for the construction of a feasible timetable. The second part evaluates the timetable-under-construction with a simulation based on linear equations. The symbiosis of these two parts leads to an optimal timetable, given the model environment. The optimization focuses on arrival delays, but other timetable characteristics such as travel time can also be incorporated in the objective function. A case-study shows that a delay reduction up to 30% is possible within the model settings. The model's strength is the improvement of existing timetables, for which the structure is unharmed during the optimization. Binary variables can be used to develop new timetables. The model can also be adapted to homogenize the timetable optimally.

To the best of our knowledge, this is the first timetabling model which takes delay propagation explicitly into account, and optimizes the timetable directly, without cumbersome trial-and-error procedures.

Chapter 7 discusses the results of the thesis and compares them with the objectives from Chapter 1. This leads to some additional research questions of which the answers can lead to additional reliability improvements of the railway traffic.

Samenvatting

Summary in Dutch

Betrouwbaarheid van spoorwegsystemen

Om de stedelijke gebieden in de toekomst tegen acceptabele kosten bereikbaar te houden, zal de benutting van de spoorweginfrastructuur moeten toenemen. Ondanks het feit dat de spoorbenutting in Nederland al een van de hoogste ter wereld is, zet de spoorsector zich in voor een betere benutting van de capaciteit. Om dit te bereiken zal het spoorverkeer wel veel betrouwbaarder moeten zijn.

Daarnaast is het op tijd rijden van de treinen zelf al de belangrijkste doelstelling van NS. Het is namelijk een belangrijke eigenschap om klanten aan zich te binden. Ook staat het op tijd rijden centraal in contracten met de overheid en reizigersorganisaties.

Het hoofddoel van het onderzoek is *"het ontwikkelen van regels en modellen waarmee het construeren van betrouwbare spoorwegdienstregelingen kan worden ondersteund"*.

Hoofdstuk 1 schetst eerst een beeld van de spoorwegwereld waarin het onderzoek plaats heeft gevonden. Daarna worden enkele begrippen beschreven, die voor het betrouwbaarheidsonderzoek van belang zijn. Zo is *de 3-minuten-aankomstpunctualiteit* het percentage treinen dat binnen drie minuten van de geplande tijd aankomt op een van de grote stations.

Voor het ontwikkelen van deze regels en modellen wordt in Hoofdstuk 2 eerst uitgelegd hoe de planning van ondermeer de dienstregeling bij een spoorbedrijf vandaag de dag geschiedt. Tijdens de uitvoering van deze dienstregeling zullen er echter verstoringen optreden door bijvoorbeeld defecten aan de infrastructuur of het materieel, of door weersomstandigheden. Door de samenhang van het spoorwegnet kunnen de daardoor ontstane vertragingen zich snel door het netwerk verspreiden. Daarom zijn in eerste instantie vele van deze afhankelijkheden binnen spoorwegsystemen beschreven. Het modeleren van spoorwegsystemen, iets wat herhaaldelijk gebeurt in latere hoofdstukken, is de kunst van het formaliseren van deze afhankelijkheden.

Hoofdstuk 3 bevat een beschrijving van een aantal belangrijke modellen voor het ontwikkelen en evalueren van dienstregelingen. Eerst komt de *DONS*-software ter sprake, dat door NS en ProRail gebruikt wordt voor het construeren van cyclische dienstregelingen. Deze cyclische dienstregeling, van 60 minuten, wordt een aantal malen achter elkaar gezet om de dienstregeling voor een dag te maken.

Met *max-plus algebra*, die alleen de operatoren 'optellen' en 'het maximum nemen' kent, kunnen voor cyclische dienstregelingen enkele karakteristieken worden berekend die betrekking hebben op de betrouwbaarheid.

Met simulatie kunnen spoorwegsystemen in detail worden nagebootst. Voor NS en ProRail is voor de beoordeling van dienstregelingen het simulatie-pakket *SIMONE* ontwikkeld.

Spoorweg- en andere openbaar vervoerbedrijven nemen speling op in hun dienstregeling om kleine vertragingen op te vangen. Er is echter slechts in beperkte mate onderzoek gedaan naar de beste allocatie van deze speling. Hoofdstuk 4 van dit proefschrift beschrijft deze problematiek en komt tot de conclusie dat de in Nederland toegepaste proportionele verdeling van speling (overal 7% van de minimale rijtijd) niet optimaal is. Het is beter minder speling in de eerste en laatste delen van treinlijnen op te nemen, en relatief wat meer in het midden. Dit volgt intuïtief uit twee oorzaken. Als treinen aan het begin van hun reis speling gebruiken, dan wordt voor bijna de hele treinlijn de vertraging verkleind. Als de speling pas aan het eind wordt gebruikt, wordt alleen de aankomst op het allerlaatste station verbeterd door deze speling. Aan de andere kant is teveel speling in het begin vaak overbodig, omdat de speling niet gebruikt kan worden als de trein nog geen vertraging heeft opgelopen. Door relatief veel speling halverwege de treinlijn te leggen kan de punctualiteit aanzienlijk worden verbeterd.

Snelheidsverschillen van verschillende treindiensten leiden, met name op drukke baanvakken, tot korte opvolgtijden. Bij korte opvolgtijden is de kans op vertragingsoverdracht tussen treinen veel groter. Door het treinverkeer te *homogeniseren*, dat wil zeggen door de snelheidsverschillen te verminderen, nemen de opvolgtijden toe en vermindert de vertragingsvoortplanting. In Hoofdstuk 5 worden deze afhankelijkheden gekwantificeerd door de twee nieuwe maatstaven SSHR en SAHR. Deze maatstaven kwantificeren niet alleen de heterogeniteit van het treinverkeer, maar geven ook een indicatie voor de vertragingsvoortplanting. Deze resultaten wordt kracht bijgezet door twee uitgebreide simulatie studies.

Het in Hoofdstuk 6 gepresenteerde model weet veel afhankelijkheden binnen spoorwegsystemen in 'n model te vangen. Het stochastisch optimalisatie model, een lineair programmeringsmodel dat afgeleid is van zogenaamde recourse modellen, bestaat uit twee delen. Het eerste deel geeft alle voorwaarden voor de constructie van een toegestane dienstregeling weer. Het tweede deel evalueert de in constructie zijnde dienstregeling met een op lineaire vergelijkingen gebaseerde simulatie. De symbiose van deze twee delen leidt tot een optimale dienstregeling, gegeven de modelomgeving. De optimalisatie vindt in eerste instantie plaats met betrekking tot de aankomstvertragingen, maar andere dienstregelingskarakteristieken zoals reistijd kunnen ook worden meegenomen in de doelstellingsfunctie. Een case-studie laten zien dat een modelmatige vertragingsvermindering tot 30% mogelijk is. Het model is vooral sterk in het verbeteren van bestaande dienstregelingen, waarbij de structuur van de dienstregeling in takt blijft. Met binaire variabelen kunnen ook nieuwe dienstregelingen worden ontwikkeld. Ook kan het model zodanig worden aangepast, dat een dienstregeling op een optimale wijze wordt gehomogeniseerd.

Voor zover wij weten is dit het eerste dienstregelingsmodel dat vertragingsvoortplanting expliciet modeleert en de dienstregeling direct optimaliseert, zonder lastige trial-and-error procedures. Bovenal verwachten we een punctualiteitsverbetering van enkele procentpunten als het model integraal wordt toegepast op een landelijke dienstregeling.

In Hoofdstuk 7 worden de resultaten van het proefschrift samengevat en vergeleken met de doelstelling uit Hoofdstuk 1. Hier vloeien een aantal vervolg-onderzoeksvragen uit voort, waarvan de beantwoording kan bijdragen tot extra verbetering van de betrouwbaarheid van het treinverkeer.

Curriculum Vitae

Michiel Vromans was born on January 31, 1976, in Veghel, the Netherlands. In 1994-1995 he visited the Emporia State University, Emporia, Kansas, USA. His focus that year was on both geography and mathematics. In 1995 he started his study in Econometrics at the University of Groningen, with a major in Operations Research. He graduated in 2000 with a Master's thesis on the automation of train crew planning. This thesis was written during a nine month internship at NS Reizigers, the largest Dutch passenger railway operator.

In October 2000, he started as a Ph.D. student at the Rotterdam School of Management, Erasmus University Rotterdam. His Ph.D. research considered the improvement of the reliability and punctuality of railway systems. Much of his research was carried out at, and in cooperation with the logistics department of NS Reizigers.

From February 2002 until August 2003, he supported NS Reizigers with two practical studies on human resource management for train personnel. This also led to an article in Interfaces.

Several other of his articles, related to this thesis, are forthcoming in scientific journals and books. Several research papers have been published conference proceedings. He has also presented his research work at various international conferences and workshops. In October 2004 the joint paper *Reliability and Heterogeneity of Railway Services* with Leo Kroon and Rommert Dekker won the first prize in the 2004 student competition of the Rail Applications Special Interest Group of INFORMS. In November 2004, he was awarded the IT&L-award of TRAIL and AVV, with a joint paper entitled *Stochastic Optimization of Railway Timetables.*

In January 2005 he joined ProRail, the Dutch railway infrastructure manager. His work there, partially in line with his dissertation research, focuses on the capacity of the Dutch railway network and timetabling norms.

Erasmus Research Institute of Management
ERIM Ph.D. Series
Research in Management

Appelman, J.H., *Governance of Global Interorganizational Tourism Networks; Changing Forms of Co-ordination between the Travel Agency and Aviation Sector*, Promotors: Prof. dr. F.M. Go & Prof. dr. B. Nooteboom, EPS-2004-036-MKT, ISBN 90-5892-060-7, http://hdl.handle.net/1765/1199

Berens, G., *Corporate Branding: The Development of Corporate Associations and their Influence on Stakeholder Reactions*, Promotor: Prof. dr. C. B. M. van Riel, EPS-2004-039-ORG, ISBN 90 -5892-065-8, http://hdl.handle.net/1765/1273

Berghe, D.A.F., *Working Across Borders: Multinational Enterprises and the Internationalization of Employment*, Promotors: Prof. dr. R.J.M. van Tulder & Prof. dr. E.J.J. Schenk, EPS-2003-029-ORG, ISBN 90-5892-05-34, http://hdl.handle.net/1765/1041

Bijman, W.J.J., *Essays on Agricultural Co-operatives; Governance Structure in Fruit and Vegetable Chains*, Promotor: Prof. dr. G.W.J. Hendrikse, EPS-2002-015-ORG, ISBN: 90-5892-024-0, http://hdl.handle.net/1765/867

Brito, M.P. de, *Managing Reverse Logistics or Reversing Logistics Management?* Promotors: Prof. dr. ir. R. Dekker & Prof. dr. M. B. M. de Koster, EPS-2004-035-LIS, ISBN: 90-5892-058-5, http://hdl.handle.net/1765/1132

Campbell, R.A.J., *Rethinking Risk in International Financial Markets*, Promotor: Prof. dr. C.G. Koedijk, EPS-2001-005-F&A, ISBN: 90-5892-008-9, http://hdl.handle.net/1765/306

Chen, Y., *Labour Flexibility in China's Companies: An Empirical Study*, Promotors: Prof. dr. A. Buitendam & Prof. dr. B. Krug, EPS-2001-006-ORG, ISBN: 90-5892-012-7, http://hdl.handle.net/1765/307

Daniševská, P., *Empirical Studies on Financial Intermediation and Corporate Policies*, Promotor: Prof. dr. C.G. Koedijk, EPS-2004-044-F&A, ISBN 90-5892-070-4, http://hdl.handle.net/1765/1518

Delporte-Vermeiren, D.J.E., *Improving the Flexibility and Profitability of ICT-enabled Business Networks: An Assessment Method and Tool*, Promotors: Prof.mr. dr. P.H.M. Vervest & Prof. dr. ir. H.W.G.M. van Heck, EPS-2003-020-LIS, ISBN:

90-5892-040-2, http://hdl.handle.net/1765/359

Dijksterhuis, M., *Organizational Dynamics of Cognition and Action in the Changing Dutch and US Banking Industries*, Promotors: Prof. dr. ing. F.A.J. van den Bosch & Prof. dr. H.W. Volberda, EPS-2003-026-STR, ISBN: 90-5892-048-8, http://hdl.handle.net/1765/1037

Fenema, P.C. van, *Coordination and Control of Globally Distributed Software Projects*, Promotor: Prof. dr. K. Kumar, EPS-2002-019-LIS, ISBN: 90-5892-030-5, http://hdl.handle.net/1765/360

Fleischmann, M., *Quantitative Models for Reverse Logistics*, Promotors: Prof. dr. ir. J.A.E.E. van Nunen & Prof. dr. ir. R. Dekker, EPS-2000-002-LIS, ISBN: 3540 417 117, http://hdl.handle.net/1765/1044

Flier, B., *Strategic Renewal of European Financial Incumbents: Coevolution of Environmental Selection, Institutional Effects, and Managerial Intentionality*, Promotors: Prof. dr. ing. F.A.J. van den Bosch & Prof. dr. H.W. Volberda, EPS-2003-033-STR, ISBN: 90-5892-055-0, http://hdl.handle.net/1765/1071

Fok, D., *Advanced Econometric Marketing Models*, Promotor: Prof. dr. P.H.B.F. Franses, EPS-2003-027-MKT, ISBN: 90-5892-049-6, http://hdl.handle.net/1765/1035

Ganzaroli , A., *Creating Trust between Local and Global Systems*, Promotors: Prof. dr. K. Kumar & Prof. dr. R.M. Lee, EPS-2002-018-LIS, ISBN: 90-5892-031-3, http://hdl.handle.net/1765/361

Gilsing, V.A., *Exploration, Exploitation and Co-evolution in Innovation Networks*, Promotors: Prof. dr. B. Nooteboom & Prof. dr. J.P.M. Groenewegen, EPS-2003-032-ORG, ISBN 90-5892-05-42, http://hdl.handle.net/1765/1040

Graaf, G. de, *Tractable Morality: Customer Discourses of Bankers, Veterinarians and Charity Workers*, Promotors: Prof. dr. F. Leijnse & Prof. dr. T. van Willigenburg, EPS-2003-031-ORG, ISBN: 90-5892-051-8, http://hdl.handle.net/1765/1038

Hermans. J.M., *ICT in Information Services, Use and deployment of the Dutch securities trade, 1860-1970.* Promotor: Prof. dr. drs. F.H.A. Janszen, EPS-2004-046-ORG, ISBN 90-5892-072-0, http://hdl.handle.net/1765/1793

Heugens, P.M.A.R., *Strategic Issues Management: Implications for Corporate Performance*, Promotors: Prof. dr. ing. F.A.J. van den Bosch & Prof. dr. C.B.M. van

Riel, EPS-2001-007-STR, ISBN: 90-5892-009-7, http://hdl.handle.net/1765/358

Hooghiemstra, R., *The Construction of Reality*, Promotors: Prof. dr. L.G. van der Tas RA & Prof. dr. A.Th.H. Pruyn, EPS-2003-025-F&A, ISBN: 90-5892-047-X, http://hdl.handle.net/1765/871

Jansen, J.J.P., *Ambidextrous Organizations*, Promotors: Prof.dr.ing. F.A.J. Van den Bosch & Prof.dr. H.W. Volberda, EPS-2005-055-STR, ISBN 90-5892-081-X, http://hdl.handle.net/1765/1

Jong, C. de, *Dealing with Derivatives: Studies on the Role, Informational Content and Pricing of Financial Derivatives*, Promotor: Prof. dr. C.G. Koedijk, EPS-2003-023-F&A, ISBN: 90-5892-043-7, http://hdl.handle.net/1765/1043

Kippers, J., *Empirical Studies on Cash Payments*, Promotor: Prof.dr. Ph.H.B.F. Franses, EPS-2004-043-F&A. ISBN 90-5892-069-0, http://hdl.handle.net/1765/1520

Koppius, O.R., *Information Architecture and Electronic Market Performance*, Promotors: Prof. dr. P.H.M. Vervest & Prof. dr. ir. H.W.G.M. van Heck, EPS-2002-013-LIS, ISBN: 90-5892-023-2, http://hdl.handle.net/1765/921

Langen, P.W. de, *The Performance of Seaport Clusters; A Framework to Analyze Cluster Performance and an Application to the Seaport Clusters of Durban, Rotterdam and the Lower Mississippi*, Promotors: Prof. dr. B. Nooteboom & Prof. drs. H.W.H. Welters, EPS-2004-034-LIS, ISBN: 90-5892-056-9, http://hdl.handle.net/1765/1133

Le Ahn, T., *Intelligent Control of Vehicle-Based Internal Transport Systems*, Promotors: Prof. dr. M.B.M. de Koster & Prof. dr. ir. R. Dekker, EPS-2005-051-LIS, ISBN: 90-5892-079-8, http://hdl.handle.net/1765/1

Liang, G., *New Competition; Foreign Direct Investment And Industrial Development In China*, Promotor: Prof. dr. R.J.M. van Tulder, EPS-2004-047-ORG, ISBN 90-5892-073-9, http://hdl.handle.net/1765/1795.

Loef, J., *Incongruity between Ads and Consumer Expectations of Advertising*, Promotors: Prof. dr. W.F. van Raaij & Prof. dr. G. Antonides, EPS-2002-017-MKT, ISBN: 90-5892-028-3, http://hdl.handle.net/1765/869

Mandele, L.M., van der, *Leadership and the Inflection Point: A Longitudinal Perspective*, Promotors: Prof. dr. H.W. Volberda, Prof. dr. H.R. Commandeur, EPS-2004-042-STR, ISBN 90-5892-067-4, http://hdl.handle.net/1765/1302

Meer, J.R. van der, *Operational Control of Internal Transport*, Promotors: Prof. dr. M.B.M. de Koster & Prof. dr. ir. R. Dekker, EPS-2000-001-LIS, ISBN:90-5892-004-6, http://hdl.handle.net/1765/859

Miltenburg, P.R., *Effects of Modular Sourcing on Manufacturing Flexibility in the Automotive Industry: A Study among German OEMs*, Promotors: Prof. dr. J. Paauwe & Prof. dr. H.R. Commandeur, EPS-2003-030-ORG, ISBN 90-5892-052-6, http://hdl.handle.net/1765/1039

Mol, M.M., *Outsourcing, Supplier-relations and Internationalisation: Global Source Strategy as a Chinese Puzzle*, Promotor: Prof. dr. R.J.M. van Tulder, EPS-2001-010-ORG, ISBN: 90-5892- 014-3, http://hdl.handle.net/1765/355

Mulder, A., *Government Dilemmas in the Private Provision of Public Goods*, Promotor: Prof. dr. R.J.M. van Tulder, EPS-2004-045-ORG, ISBN: 90-5892- 071-2, http://hdl.handle.net/1765/1790

Muller, A.R., *The Rise of Regionalism: Core Company Strategies Under The Second Wave of Integration*, Promotor: Prof. dr. R.J.M. van Tulder, EPS-2004-038-ORG, ISBN 90-5892-062-3, http://hdl.handle.net/1765/1272

Oosterhout, J. van, *The Quest for Legitimacy: On Authority and Responsibility in Governance*, Promotors: Prof. dr. T. van Willigenburg & Prof.mr. H.R. van Gunsteren, EPS-2002-012-ORG, ISBN: 90-5892-022-4, http://hdl.handle.net/1765/362

Peeters, L.W.P., *Cyclic Railway Timetable Optimization*, Promotors: Prof. dr. L.G. Kroon & Prof. dr. ir. J.A.E.E. van Nunen, EPS-2003-022-LIS, ISBN: 90-5892-042-9, http://hdl.handle.net/1765/429

Popova. V., *Knowledge Discovery and Monotonicity*, Promotor: Prof. dr. A. de Bruin, EPS-2004-037-LIS, ISBN 90-5892-061-5, http://hdl.handle.net/1765/1201

Puvanasvari Ratnasingam, P., *Interorganizational Trust in Business to Business E-Commerce*, Promotors: Prof. dr. K. Kumar & Prof. dr. H.G. van Dissel, EPS-2001-009-LIS, ISBN: 90-5892-017-8, http://hdl.handle.net/1765/356

Romero Morales, D., *Optimization Problems in Supply Chain Management*, Promotors: Prof. dr. ir. J.A.E.E. van Nunen & Dr. H.E. Romeijn, EPS-2000-003-LIS, ISBN: 90-9014078-6, http://hdl.handle.net/1765/865

Roodbergen , K.J., *Layout and Routing Methods for Warehouses*, Promotors: Prof. dr. M.B.M. de Koster & Prof. dr. ir. J.A.E.E. van Nunen, EPS-2001-004-LIS, ISBN:

90-5892-005-4, http://hdl.handle.net/1765/861

Schweizer, T.S., *An Individual Psychology of Novelty-Seeking, Creativity and Innovation*, Promotor: Prof. dr. R.J.M. van Tulder, EPS-2004-048-ORG, ISBN 90-5892-07-71, http://hdl.handle.net/1765/1271

Six, F.E., *Trust and Trouble: Building Interpersonal Trust Within Organizations*, Promotors: Prof. dr. B. Nooteboom & Prof. dr. A.M. Sorge, EPS-2004-040-ORG, ISBN 90-5892-064-X, http://hdl.handle.net/1765/1271

Slager, A.M.H., *Banking across Borders*, Promotors: Prof. dr. D.M.N. van Wensveen & Prof. dr. R.J.M. van Tulder, EPS-2004-041-ORG, ISBN 90-5892-066-6, http://hdl.handle.net/1765/1301

Speklé, R.F., *Beyond Generics: A closer look at Hybrid and Hierarchical Governance*, Promotor: Prof. dr. M.A. van Hoepen RA, EPS-2001-008-F&A, ISBN: 90-5892-011-9, http://hdl.handle.net/1765/357

Teunter, L.H., *Analysis of Sales Promotion Effects on Household Purchase Behavior*, Promotors: Prof. dr. ir. B. Wierenga & Prof. dr. T. Kloek, EPS-2002-016-ORG, ISBN: 90-5892-029-1, http://hdl.handle.net/1765/868

Valck, K. de, *Virtual Communities of Consumption: Networks of Consumer Knowledge and Companionship*, Promotors: Prof.dr.ir. G.H. van Bruggen, & Prof.dr.ir. B. Wierenga, EPS-2005-050-MKT, ISBN 90-5892-078-X, http://hdl.handle.net/1765/861

Verheul, I., *Is there a (fe)male approach? Understanding gender differences in entrepreneurship*, Prof.dr. A.R. Thurik, EPS-2005-054-ORG, ISBN 90-5892-080-1, http://hdl.handle.net/1765/1

Vis, I.F.A., *Planning and Control Concepts for Material Handling Systems*, Promotors: Prof. dr. M.B.M. de Koster & Prof. dr. ir. R. Dekker, EPS-2002-014-LIS, ISBN: 90-5892-021-6, http://hdl.handle.net/1765/866

Vliet, P. van, *Downside Risk and Empirical Asset Pricing*, Promotor: Prof. dr. G.T. Post, EPS-2004-049-F&A, ISBN: 90-5892-07-55, http://hdl.handle.net/1765/870

Waal, T. de, *Processing of Erroneous and Unsafe Data*, Promotor: Prof. dr. ir. R. Dekker, EPS-2003-024-LIS, ISBN: 90-5892-045-3, http://hdl.handle.net/1765/870

Wielemaker, M.W., *Managing Initiatives: A Synthesis of the Conditioning* *and Knowledge-Creating View*, Promotors: Prof. dr. H.W. Volberda & Prof. dr. C.W.F. Baden-Fuller, EPS-2003-28-STR, ISBN 90-5892-050-X, http://hdl.handle.net/1765/1036

Wijk, R.A.J.L. van, *Organizing Knowledge in Internal Networks: A Multilevel Study*, Promotor: Prof. dr. ing. F.A.J. van den Bosch, EPS-2003-021-STR, ISBN: 90-5892-039-9, http://hdl.handle.net/1765/347

Wolters, M.J.J., *The Business of Modularity and the Modularity of Business*, Promotors: Prof. mr. dr. P.H.M. Vervest & Prof. dr. ir. H.W.G.M. van Heck, EPS-2002-011-LIS, ISBN: 90-5892-020-8, http://hdl.handle.net/1765/920

Reliability of Railway Systems

Railway transport plays a key role in mobility in the Netherlands and other countries. It has been recognized that the on-time performance is one of the key performance indicators in railway transport. Many different internal and external factors cause the train operations to be disturbed. Moreover, incurred delays are often propagated to other trains and to other parts of the network. The societal, managerial, and scientific relevance of research on the on-time performance of railway systems are eminent. This thesis provides a clear picture of the reliability of railway systems. A railway system can be considered as a very large and complex stochastic dynamic system. "Reliability of Railway Systems" describes mathematical models for the evaluation and optimization of railway timetables. Special attention is given to the allocation of running time supplements. These supplements can be very useful in containing delay propagation. However, the effectiveness of these supplements highly depends on the location within a train line. A surprising, but potentially effective supplement allocation rule is developed to decrease the propagation of delays. Another important subject is the heterogeneity of train traffic, or in other words the speed differences. Besides showing a strong correlation between speed differences and reliability, new measures were developed to capture the heterogeneity. Furthermore, an innovative stochastic linear program is presented that is not only able to evaluate, but also to optimize timetables. It integrates most railway dependencies, and it directly optimizes the average arrival delays. The model shows that considerable improvements are possible with respect to the current timetable in the Netherlands. Several variants of the model are described, such that the model can be used for a wide range of problems.

ERIM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are RSM Erasmus University and the Erasmus School of Economics. ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focussed on the management of the firm in its environment, its intra- and inter-firm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced graduate program in Research in Management. Within ERIM, over two hundred senior researchers and Ph.D. candidates are active in the different research programs. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

www.erim.eur.nl

ISBN 90-5892-089-5