Congested railways

Influence of infrastructure and timetable properties on delay propagation

ANDERS LINDFELDT

Licentiate Thesis in Transport Science
Stockholm, Sweden 2012
Abstract
In this thesis the symptoms and underlying behaviour of congestion on railways are analysed and discussed. As well as in many other countries, Sweden faces increasing demand for transport. To meet this new demand, railways play an important role. Today, the capacity of the Swedish rail network is not upgraded at the speed necessary to keep up with the increase in traffic demand. The sensitivity of the railway system rises as the capacity utilisation increases. At some point the marginal gain of operating one extra train is lower than the costs in term of increased sensitivity to delay, i.e. maximum capacity has been reached.

Two methodologies are employed in this thesis to analyse capacity. The first uses real data from the Swedish rail network, train operation and delays to analyse how different factors influence available capacity and delay creation. Several useful key performance indicators are defined to describe capacity influencing properties of the infrastructure and the rail traffic. The rail network is divided into subsections for which the indicators have been estimated. This makes it possible to discern their different characteristics and identify potential weaknesses.

The second approach employs the railway simulation tool RailSys in extensive simulation experiments. This methodology is used to analyse the characteristics of double track operation. Simulation of several hundred scenarios are conducted to analyse the influence of traffic density, timetable speed heterogeneity, primary delays and inter-station distance on secondary delays and used timetable allowance. The analysis gives an in-depth understanding of the mechanisms behind the performance of a double track.
Acknowledgements

Trafikverket (Swedish Transport Administration) provided funding for this research.

I would like to thank my supervisors Bo-Lennart Nelldal and Lars-Göran Mattsson for their support and advice. Thanks also to Magnus Wahlborg, a member of the reference group and contact person at Trafikverket responsible for the research program financing this thesis. Thanks to Magdalena Grimm at Trafikverket, also a member of the reference group.

Thanks to Hans Sipilä, Jennifer Warg and Olov Lindfeldt for all the stimulating discussions and their valuable inputs.

Finally, thanks to all colleagues and fellow PhD students at ToL for making work fun and joining me for the many and too long coffee breaks.

Stockholm, May 2012

Anders Lindfeldt
List of Publications

Reports


Papers


### Contents

1. Introduction .......................................................................................................................... 1
   1.1 Background .................................................................................................................. 1
   1.2 Objectives .................................................................................................................... 2
   1.3 Delimitations ............................................................................................................... 3
   1.4 Structure of the thesis ................................................................................................. 3
2. Related research ..................................................................................................................... 5
3. Concepts .................................................................................................................................... 9
4. Methodology .......................................................................................................................... 13
   4.1 Performance evaluation of the Swedish railway network ......................................... 13
   4.2 Simulation analysis of double track operation .......................................................... 17
5. Results ...................................................................................................................................... 23
   5.1 Performance evaluation of the Swedish railway network ......................................... 23
   5.2 Simulation analysis of double track operation .......................................................... 28
6. Conclusions ............................................................................................................................ 33
7. Contribution of the thesis .................................................................................................... 37
8. Future research ....................................................................................................................... 39
9. References ............................................................................................................................ 43
1 Introduction

1.1 Background

The demand for transportation on railways grows for each year and many railway lines around the greater cities in Sweden are already used close to maximum capacity, especially during rush hours. In order to meet the increasing demand, either new railway lines have to be built, or the existing ones need to be upgraded and used more efficiently. Building new lines is associated with high costs and requires long and careful planning before construction can begin. Normally in the cities where the need is greatest, the cost and time for construction are also the highest. For this reason, upgrading and increasing the efficiency of already existing lines is an attractive option. Figure 1.1 shows the increase in traffic in Sweden during the last years.

Figure 1.1: Increase or decrease in trains/day in Sweden compared to 2008 for year 2009-2011. The range is 50-150% where green means decrease, yellow no change and red increase. Only line sections with more than 10 trains/day are shown in colour.
In order to take the eight actions to increase capacity, we have to understand how the railway system works and respond to increased capacity utilisation. Railway operation involves several complex systems, such as railway infrastructure, rolling stock, timetable, human behaviour etc. The challenge of analysing and describing railway operation lies in its complexity and the highly nonlinear behaviour. Some of the factors influencing capacity are: track layout, speed difference between train services, market demand, level of primary delays and acceptance of delays. For this reason, there is not an easy way to tell what the maximum capacity of a railway line is, unless all parameters are specified in detail. But then the measure lose some of its generality and usefulness.

Utilizing a railway optimally is often a question about quantity versus quality, i.e. no. trains vs. delays. Increasing no. trains leads to a higher sensitivity to delays with more secondary delays propagating from train to train. Quality can also include the quality of the scheduled train path with regard to longer travel times due to lower mean speed. To define capacity, these factors have to be weighed against each other to find the best solution. There is a conflict of interest between adding new train paths to meet higher demand and maintaining the quality of the already scheduled trains. With the upcoming deregulation of operation in Sweden, the necessity to be able to solve this conflict accurately increases when service operators that are denied train slots due to congestion demand a valid motivation.

This licentiate thesis is part of the research project Congested railways that is included in a research program for capacity analysis and simulation at KTH, with the aim to develop and improve methods in this field. The program is developed in cooperation with Trafikverket (the Swedish Transport Administration) who also provides funding for most of the research done in the capacity field.

1.2 Objectives

The overall objective of this thesis is to analyse what happens when the traffic intensity on a railway line approaches maximum capacity and to improve the understanding about the underlying behaviour of a congested railway system. What are the symptoms, and how do different infrastructure and timetable properties affect maximum capacity? Capacity is not easily defined and one objective of this thesis is to provide insight into how it can be done and find parameters of importance. The capacity of a railway line can then be expressed as a function of these parameters. The infrastructure may be considered constant in the capacity evaluation, or represented by a few alternatives, but the results should not be timetable specific.

To reach the objectives mentioned above, the first step is to identify suitable parameters affecting capacity. The second step is to establish how they affect capacity. Methods for analysis of empirical data from train operation on the Swedish rail network need to be developed, as well as the methodology of using RailSys, a timetable planning and micro simulation software for railway operation Radtke (2005), in capacity analysis.

In order to take the eight actions to increase capacity, we have to understand how the railway system works and respond to increased capacity utilisation. Railway operation involves several complex systems, such as railway infrastructure, rolling stock, timetable, human behaviour etc. The challenge of analysing and describing railway operation lies in its complexity and the highly nonlinear behaviour. Some of the factors influencing capacity are: track layout, speed difference between train services, market demand, level of primary delays and acceptance of delays. For this reason, there is not an easy way to tell what the maximum capacity of a railway line is, unless all parameters are specified in detail. But then the measure lose some of its generality and usefulness.

Utilizing a railway optimally is often a question about quantity versus quality, i.e. no. trains vs. delays. Increasing no. trains leads to a higher sensitivity to delays with more secondary delays propagating from train to train. Quality can also include the quality of the scheduled train path with regard to longer travel times due to lower mean speed. To define capacity, these factors have to be weighed against each other to find the best solution. There is a conflict of interest between adding new train paths to meet higher demand and maintaining the quality of the already scheduled trains. With the upcoming deregulation of operation in Sweden, the necessity to be able to solve this conflict accurately increases when service operators that are denied train slots due to congestion demand a valid motivation.

This licentiate thesis is part of the research project Congested railways that is included in a research program for capacity analysis and simulation at KTH, with the aim to develop and improve methods in this field. The program is developed in cooperation with Trafikverket (the Swedish Transport Administration) who also provides funding for most of the research done in the capacity field.

1.2 Objectives

The overall objective of this thesis is to analyse what happens when the traffic intensity on a railway line approaches maximum capacity and to improve the understanding about the underlying behaviour of a congested railway system. What are the symptoms, and how do different infrastructure and timetable properties affect maximum capacity? Capacity is not easily defined and one objective of this thesis is to provide insight into how it can be done and find parameters of importance. The capacity of a railway line can then be expressed as a function of these parameters. The infrastructure may be considered constant in the capacity evaluation, or represented by a few alternatives, but the results should not be timetable specific.

To reach the objectives mentioned above, the first step is to identify suitable parameters affecting capacity. The second step is to establish how they affect capacity. Methods for analysis of empirical data from train operation on the Swedish rail network need to be developed, as well as the methodology of using RailSys, a timetable planning and micro simulation software for railway operation Radtke (2005), in capacity analysis.
1.3 Delimitations

The capacity of a railway can have many definitions. In this thesis capacity is referred to as the number of trains/h that can be operated on a railway line. Consequently it does not include parameters like for example train size or limitations in availability of the infrastructure due to maintenance. Also when referring to capacity, it is neither that of specific stations or short line sections, nor that of a large scale network, but rather that of a longer railway line.

The economic effects of high capacity utilisation are in this work limited to scheduled and operational delays. Even though delays are closely connected to socioeconomic costs and direct costs for the operators, they are not considered in this thesis. Neither is the investment cost for building new infrastructure or increasing maintenance to reduce primary delays or the market perspective of service frequency, delays and not being able to run more trains due to congestion.

The analysis performed in this thesis address railway traffic during normal operation, which means operation without too large delays, i.e. interruptions. In real life this can typically be situations resulting in longer times of complete stops, like vehicle or catenary failure on a line section. In situations like these trains are cancelled and redirected to use other routes in the railway network. Consequently, they are hard to analyse and require other methods of analysis than employed in this thesis.

1.4 Structure of the thesis

Section 2 gives an overview of some relevant studies regarding simulation of train traffic, timetable properties and possible evaluation parameters. An introduction to capacity related concepts are given in section 3. Section 4 discusses the methodologies employed in this thesis. Section 5 covers the results from the report and two papers that make up the second part of this thesis. Sections 6-8 consist of conclusions, contribution of the thesis and future research.
2 Related research

There are several different methods of analysing railway operation. The different methods can be divided into analytical, combinatorial and simulation based, Mattson (2007). All approaches have their advantages and disadvantages. Typically, the advantages of analytical and combinatorial methods are that they do not necessarily require detailed information about for example the timetable. This makes them suitable to long-term planning where a timetable may not exist and general results are needed. Among the disadvantages are that perturbations often are modelled in a simplified manner, if they are modelled at all, and that the effect of dispatching is not considered. Simulation on the other hand can model the perturbations in detail, but is in general time consuming and requires detailed knowledge about timetable and infrastructure.

Huisman (1999) developed a stochastic model for estimating the running time on double track railway lines with heterogeneous train traffic. The model describes secondary delays due to faster trains catching up with slower ones. The train order can be either random, which is useful for long term planning, or defined by a cyclic timetable. The primary delays used include both entry delays and running time extensions. Huisman demonstrates the model by applying it on a Dutch railway line to show how the number of trains, heterogeneity, primary delay, train order and buffer times influence the delays. However, the model is limited to analyse delays on line sections where trains are not allowed to overtake, hence delays at stations due to overtaking and dispatching actions are not included.

Vromans (2006) defines two measures of heterogeneity and uses simulation to show their correlation to the average delay. The two measures are SSHR (sum of shortest headway reciprocals) and SAHR (sum of arrival headway reciprocals). The first measure looks at the headway both at the start and at the end of the line section, and therefore takes into consideration both the heterogeneity in speed of the trains and the spread of the trains over time. The second measure, SAHR, focus only at the headway at the end of the line section under the assumption that the headway at the end is more important than at the start. Several timetables with different heterogeneity are created and simulated using the simulation tool SIMONE to show that both heterogeneity measures correlate positively to the average delay. In the simulation both dwell time extensions and running time extensions are used. Overtakings are also possible.

Gorman (2009) uses real data to do statistical estimations of delays. He predicts total train running time based on free running time predictors and congestion-related factors, such as meets, passes, overtakes, train spacing variability and departure headway. He concludes that the factors showing largest effect on congestion delay are meets, passes and overtakes.

Rudolph (2003) looks into the effect of size and allocation of different types of timetable margins and allowances. Higher allowances and margins increase timetable stability but also increase travel times and reduce capacity. Allowances are defined as additional time added to the technical minimum running time between the stations or time added to the stop time at the station. The allowance is meant to compensate for small delays to avoid that the train gets late. The margin, or buffer time, is added to the minimum headway between two trains with the purpose to avoid delay transfer between trains. Different strategies for applying allowance and margins are investigated using the simulation tool RailSys.

Rudolph also categorizes delays by cause and effect. By cause, the delays are divided into primary and secondary delays. The primary delays are created with no other trains involved faster trains catching up with slower ones. The train order can be either random, which is useful for long term planning, or defined by a cyclic timetable. The primary delays used include both entry delays and running time extensions. Huisman demonstrates the model by applying it on a Dutch railway line to show how the number of trains, heterogeneity, primary delay, train order and buffer times influence the delays. However, the model is limited to analyse delays on line sections where trains are not allowed to overtake, hence delays at stations due to overtaking and dispatching actions are not included.

Vromans (2006) defines two measures of heterogeneity and uses simulation to show their correlation to the average delay. The two measures are SSHR (sum of shortest headway reciprocals) and SAHR (sum of arrival headway reciprocals). The first measure looks at the headway both at the start and at the end of the line section, and therefore takes into consideration both the heterogeneity in speed of the trains and the spread of the trains over time. The second measure, SAHR, focus only at the headway at the end of the line section under the assumption that the headway at the end is more important than at the start. Several timetables with different heterogeneity are created and simulated using the simulation tool SIMONE to show that both heterogeneity measures correlate positively to the average delay. In the simulation both dwell time extensions and running time extensions are used. Overtakings are also possible.

Gorman (2009) uses real data to do statistical estimations of delays. He predicts total train running time based on free running time predictors and congestion-related factors, such as meets, passes, overtakes, train spacing variability and departure headway. He concludes that the factors showing largest effect on congestion delay are meets, passes and overtakes.

Rudolph (2003) looks into the effect of size and allocation of different types of timetable margins and allowances. Higher allowances and margins increase timetable stability but also increase travel times and reduce capacity. Allowances are defined as additional time added to the technical minimum running time between the stations or time added to the stop time at the station. The allowance is meant to compensate for small delays to avoid that the train gets late. The margin, or buffer time, is added to the minimum headway between two trains with the purpose to avoid delay transfer between trains. Different strategies for applying allowance and margins are investigated using the simulation tool RailSys.

Rudolph also categorizes delays by cause and effect. By cause, the delays are divided into primary and secondary delays. The primary delays are created with no other trains involved faster trains catching up with slower ones. The train order can be either random, which is useful for long term planning, or defined by a cyclic timetable. The primary delays used include both entry delays and running time extensions. Huisman demonstrates the model by applying it on a Dutch railway line to show how the number of trains, heterogeneity, primary delay, train order and buffer times influence the delays. However, the model is limited to analyse delays on line sections where trains are not allowed to overtake, hence delays at stations due to overtaking and dispatching actions are not included.

Vromans (2006) defines two measures of heterogeneity and uses simulation to show their correlation to the average delay. The two measures are SSHR (sum of shortest headway reciprocals) and SAHR (sum of arrival headway reciprocals). The first measure looks at the headway both at the start and at the end of the line section, and therefore takes into consideration both the heterogeneity in speed of the trains and the spread of the trains over time. The second measure, SAHR, focus only at the headway at the end of the line section under the assumption that the headway at the end is more important than at the start. Several timetables with different heterogeneity are created and simulated using the simulation tool SIMONE to show that both heterogeneity measures correlate positively to the average delay. In the simulation both dwell time extensions and running time extensions are used. Overtakings are also possible.

Gorman (2009) uses real data to do statistical estimations of delays. He predicts total train running time based on free running time predictors and congestion-related factors, such as meets, passes, overtakes, train spacing variability and departure headway. He concludes that the factors showing largest effect on congestion delay are meets, passes and overtakes.

Rudolph (2003) looks into the effect of size and allocation of different types of timetable margins and allowances. Higher allowances and margins increase timetable stability but also increase travel times and reduce capacity. Allowances are defined as additional time added to the technical minimum running time between the stations or time added to the stop time at the station. The allowance is meant to compensate for small delays to avoid that the train gets late. The margin, or buffer time, is added to the minimum headway between two trains with the purpose to avoid delay transfer between trains. Different strategies for applying allowance and margins are investigated using the simulation tool RailSys.

Rudolph also categorizes delays by cause and effect. By cause, the delays are divided into primary and secondary delays. The primary delays are created with no other trains involved faster trains catching up with slower ones. The train order can be either random, which is useful for long term planning, or defined by a cyclic timetable. The primary delays used include both entry delays and running time extensions. Huisman demonstrates the model by applying it on a Dutch railway line to show how the number of trains, heterogeneity, primary delay, train order and buffer times influence the delays. However, the model is limited to analyse delays on line sections where trains are not allowed to overtake, hence delays at stations due to overtaking and dispatching actions are not included.
and can be caused by either internal or external reasons. Examples of internal reasons are
technical failures, engineering work and delays caused by railway personnel. External reasons
are extended boarding and alighting times, accidents and weather conditions. Secondary
delays are caused by interaction with other trains and are due to occupation conflicts, e.g.
headway and crossing conflicts, or delay transfer between trains with a scheduled connection.
The effects of the delays are running time extensions, dwell time extensions and additional
stops.

Lindfeldt (2010) uses advanced experimental design, simulation and response surface
metamodelling to analyse how nine different parameters affect delay development of mixed
traffic on a single track freight network. The analysis is performed with a micro simulation
software called Rail Traffic Controller, RTC, and the measure of performance is delay of the freight
trains in min per 100 train miles. The delay includes both times for meets and passes, i.e. they are not planned in
advance, and are calculated by RTC. Traffic density is varied and heterogeneity is controlled
by systematically adding passenger trains of different speeds. For completely homogeneous
freight traffic, delays are found to increase exponentially with traffic density. A relationship
between speed difference between trains and delays of the slower trains is proposed. At higher
traffic densities, the delays of freight trains increase with speed difference, but with high
enough speed difference, the effect diminishes.

In the UIC code 406 (2004), a methodology for capacity evaluation is developed with the
intention of creating an international standard. The fundamental feature of the proposed
method is to compress an existing timetable, in time, to calculate the infrastructure occupation
time. By adding buffer times to the infrastructure occupation time timetable stability is
ensured and capacity consumption achieved. The analysed railway line is divided into
appropriate sections for which the compression is made separately. The sections are chosen to
match changes in traffic patterns and infrastructure standard. If the compression indicates free
capacity, it should be attempted to enter a new train path to determine if the unused capacity
can be used or if it is to be considered lost. It is emphasised that capacity depends on the type
of traffic on the line, and that the results only are valid for the specific timetable analysed.
However, by analysing an existing timetable, market needs are considered.

capacity. He states that one problem is that many important parameters in the method are not
explained clearly and it is up to the user to decide what values to apply. Another question
raised is how the enrichment process, where additional train paths are entered into the
and can be caused by either internal or external reasons. Examples of internal reasons are
technical failures, engineering work and delays caused by railway personnel. External reasons
are extended boarding and alighting times, accidents and weather conditions. Secondary
delays are caused by interaction with other trains and are due to occupation conflicts, e.g.
headway and crossing conflicts, or delay transfer between trains with a scheduled connection.
The effects of the delays are running time extensions, dwell time extensions and additional
stops.

Lindfeldt (2010) uses advanced experimental design, simulation and response surface
metamodelling to analyse how nine different parameters affect delay development of mixed
traffic on a single track freight network. The analysis is performed with a micro simulation
software called Rail Traffic Controller, RTC, and the measure of performance is delay of the freight
trains in min per 100 train miles. The delay includes both times for meets and passes, i.e. they are not planned in
advance, and are calculated by RTC. Traffic density is varied and heterogeneity is controlled
by systematically adding passenger trains of different speeds. For completely homogeneous
freight traffic, delays are found to increase exponentially with traffic density. A relationship
between speed difference between trains and delays of the slower trains is proposed. At higher
traffic densities, the delays of freight trains increase with speed difference, but with high
enough speed difference, the effect diminishes.

In the UIC code 406 (2004), a methodology for capacity evaluation is developed with the
intention of creating an international standard. The fundamental feature of the proposed
method is to compress an existing timetable, in time, to calculate the infrastructure occupation
time. By adding buffer times to the infrastructure occupation time timetable stability is
ensured and capacity consumption achieved. The analysed railway line is divided into
appropriate sections for which the compression is made separately. The sections are chosen to
match changes in traffic patterns and infrastructure standard. If the compression indicates free
capacity, it should be attempted to enter a new train path to determine if the unused capacity
can be used or if it is to be considered lost. It is emphasised that capacity depends on the type
of traffic on the line, and that the results only are valid for the specific timetable analysed.
However, by analysing an existing timetable, market needs are considered.

capacity. He states that one problem is that many important parameters in the method are not
explained clearly and it is up to the user to decide what values to apply. Another question
raised is how the enrichment process, where additional train paths are entered into the
timetable if possible, should be done. More train paths can be added if paths are allowed to be bundled, which however might not be acceptable from a market perspective. One of the main conclusions is also that the method needs to be completed with capacity calculations of station areas, as experience has showed that this sometimes is the limiting factor rather than the line capacity. However, some of these issues will be addressed in the 2nd version of the UIC 406 capacity leaflet that is expected to be approved by UIC during 2012.

Magnarini (2010) uses several different methods of capacity evaluation to investigate the effect of different signalling systems on the capacity of a double track bottleneck in Stockholm. The methods employed are the UIC code 406, the so called Streele-formula and micro simulation. The study shows the effect of the methods having different approaches to assure timetable stability. In case of the UIC method, appropriate buffer times between trains are calculated according to a recommended maximum limit of infrastructure occupation, and do consequently not consider the shape and size of the delay distribution in the specific case study. The Streele-formula calculates the needed buffer times based on, among other parameters, accepted unscheduled waiting times, average delay at entry and probability of delay at entry of the line section, hence the delay distribution has effect on the results. The third method employs stochastic micro simulation using the software RailSys. Besides showing the difference of the three methods, the work also shows that upgrading the current Swedish signalling system, ATC2, to ERTMS Level 2 gives almost no increase in capacity of the bottleneck. ERTMS Level 3, utilising moving block, is shown to have a positive impact on capacity. The results are summarized in figure 2.1.

![Figure 2.1: Estimated capacity (trains/h) of the bottleneck in Stockholm using different methods. Magnarini (2010).](image)

Much of the research focuses on either scheduling and timetable optimisation or on explaining the creation of train delays during very specific conditions. This thesis aims to explain the behaviour of train operation on a whole railway line during different operating conditions. This includes trains overtaking and crossing, and implies that dispatching functionality has to be included in the analysis. In addition, the developed methods and results should be applicable in analyses of real rail networks. Therefore, analysis of empirical data and simulation are attractive tools. Even though there exists studies of empirical data using advanced statistical tools, it can still be hard to quantify the effect of the parameters if interest. Consequently, the analysis of empirical data conducted in this thesis is completed with a simulation experiment. Many of the simulation studies focus on explaining the impact of just one or two parameters on for example on time performance. However, there are few simulation studies to be found that explains capacity without being restricted to one or very few timetable scenarios.

![Figure 2.1: Estimated capacity (trains/h) of the bottleneck in Stockholm using different methods. Magnarini (2010).](image)
3 Concepts

The most important factor for capacity is the number of tracks on the line. The most common configurations are single-, double-, and quadruple tracks. In general the capacity of a double track is four times that of a single track, and a quadruple track three times that of a double track given a fairly heterogeneous traffic. Going from single track to double track means that trains can meet everywhere on the line without being restricted to do this only at crossing stations. Besides increasing the capacity, traffic in different direction becomes almost independent, i.e. less risk of delay transfer. On quadruple tracks, trains going in the same direction can preferably be separated according to mean speed, and is the reason why the potential capacity of a quadruple track is more than two double tracks.

Signalling combined with track layout can be crucial to capacity. For a conventional signalling system with fixed block sections, the length of the block sections on the line is of importance for the minimum headway between two consecutive trains in the same direction. Shorter block sections give shorter minimum headway times, figure 3.1, and given a limited number of block sections on a line section, they should be designed so that they have as equal occupancy time as possible. This implies that the block sections should be shorter where trains are moving slower, e.g. around and at stopping locations. For single track lines, short inter-station distances and simultaneous entry capability to decreases crossing time are crucial.

In this work, station is used for points in the network where overtaking, crossing or direction reversal is possible. Line sections are the sections of track between the stations. Distance between crossing/siding stations affects capacity in a similar way as the speed of the trains. Shorter distances mean that crossings and overtakings can be performed more often and more trains can be scheduled. For a given traffic density, more frequent siding/crossing possibilities also decreases need for scheduled delay.

Trains are normally operated according to a timetable. The timetable is needed for the passengers utilising the train services. Another fundamental reason is that trains only can meet or pass each other at discrete locations. As a consequence several aspects of the timetable must be considered when capacity analysed. The performance of the timetable decides the number of trains that can be scheduled and their scheduled running time, but it also has a direct impact on the robustness of the system, i.e. how easily delays propagate between trains. When a timetables is constructed all of these aspects, and many more, have to be considered, which is a non-trivial task.

Heterogeneity can be used to describe two different properties of the timetable. The first one is how evenly distributed the train movements are over a given period of time. The second one is associated with the speed variations between trains in the timetable. In heterogeneous timetables, trains are using the infrastructure unevenly over time with great difference in average speed. A high heterogeneity increases the risk for delay transfer, i.e secondary delays. In the first case, the buffer times between trains are unnecessary small and in the second case the speed difference implies that faster trains risk catching up on slower trains and slower trains are forced to stand aside for unscheduled overtakings. Heterogeneity due to speed differences may also introduce more allowances, for the slower trains as extra stops due to overtakings and for faster trains as extra running time allowance due to speed homogenisation. This time is called scheduled delay.
Train speed becomes an important factor for capacity especially on single track lines, where higher average speed means that crossings can occur closer in time, Nelldal (2009). On conventional double track lines speed does not have as big impact on capacity as on single track lines, even though there is a similar effect for the frequency of overtakings as for crossings on the single track line. However, if the signalling is based on fixed signal block sections, higher speeds mean that the block sections are cleared faster. The effect is somewhat counteracted by trains having longer breaking distances at higher speeds, see figure 3.1. Consequently the performance of the rolling stock in terms of acceleration and breaking performance is an important factor. It does not only mean higher mean speeds and shorter breaking distances, it does also decrease the sensitivity to delays thanks to smaller time losses due to unplanned stops or speed restrictions.

![Figure 3.1: Example of minimum technical headway on a double track section equipped with Swedish ATC2 with infill, Lindfeldt (2008).](image)

Trains stops reduce average speed and can in many cases be the major source for heterogeneity on a double track. A common example is a line where a local passenger service with frequent stops shares a track with long distance trains. Even though the top speed of the local trains may not be that much lower than the long distance trains, the frequent stopping lowers the mean speed of the local trains considerably. Stops are also important for the creation and reduction of delays. Passenger or goods exchange may take longer time than scheduled. However, if the train arrives late and is able to perform a shorter stop than scheduled, the delay will be reduced. This is especially the case for longer stops scheduled due to overtakings or crossings. Figure 3.2 shows empirical data of arrival and departure lateness at scheduled stops. Stops have taken longer than plan if the observation is above the red line and shorter if it is below it.

![Figure 3.2: Empirical data of arrival and departure lateness at scheduled stops.](image)
Allowance is extra time in the timetable that is added to the scheduled time of the trains. It can both be used to extend the running time between stations, running time allowance, or to make longer stops, allowance at stations. In both cases, the allowance can be used by the train to recover from suffered delays. The allowance may increase stability, but longer scheduled running times are negative from a market perspective. It is common to apply allowances before large junctions in the network to compensate for interference with other train movements (including shunting) and before the last station to improve the punctuality at the terminus.

Buffer time is the time between trains in the timetable. Larger buffer times reduce the probability of delay transfer between trains but also decrease the capacity. The amount of buffer time needed between trains depends on signalling system, infrastructure layout and expected severity of the delays. Often, minimum values for buffer times in different situations are used in the timetable construction, e.g. at crossings and overtakings.

On a general level, delays can be categorized into two different groups: primary delays and secondary delays. Primary delays are delays caused by faults in technical systems, human behaviour or other external factors such as severe weather conditions. Examples of sources of primary delays are faults on switches, signalling and rolling stock or stops taking longer time than planned. Primary delays can be influenced by choice of technology, education of personnel, weather conditions, wear and maintenance of infrastructure and rolling stock.

A secondary delay occurs when the source of the delay is another train. The most common reason for this delay transfer is that several trains need for same resource at the same time and thus one have to wait for it to get free. Such resources can be signal block sections, switches or platform tracks at stations. A source for secondary delay that is not due to lack of resources is when a connecting train gets delayed because it awaits the late arrival of another train.

When an isolated part of a bigger train network is analysed, two additional types of delays need to be defined: entry delay and exit delay. That means the delay the trains has when it enters and leaves the analysed system.
4 Methodology

The combination of some of the properties discussed in the previous section can be used to make a general model that explains how trains passing through an investigated network react to exogenous delays. The exit delay of a train running from the origin to the destination depends on the departure lateness at the origin (entry delay), possible primary and secondary delays suffered from along the way and any allowances in the timetable that can be used to reduce delay. The amount of allowance available in the timetable depends both on specified requirements in the timetable construction and on other factors such as capacity utilisation and priority rules. Finding the relationships of the equation below is not an easy task, especially if the results are intended to be general. Just the variables timetable and infra below require several parameters to be described, even in a simplified manner.

\[ d_e = d_a + d_p + d_s - a_u \]

Several different approaches of analysis exist and involve simulation, optimisation, queue theory and other analytical methods. All methods have their specific strengths and weaknesses and use models with different levels of detail. In general, methods based on less detailed models may be better for drawing general conclusions, which make them suitable tools for long term planning. On the other hand, more detailed models are required to perform thorough studies, but they do also require more data as input and risk generating results that are only valid for a specific setup. This thesis employs both evaluation of empirical data and micro simulation.

4.1 Performance evaluation of the Swedish railway network

Two common reasons for initialising an analysis of a railway network are that either something is not working satisfactory or some prediction needs to be made about the future. If the analysis is about a network currently in use, analysing real operational data may be the first option, rather than using a model. Such an analysis has both advantages and disadvantages. One of the biggest advantages is that if the data comes from a real system and is not the output from a model, it cannot be questioned. Of course, requires that the data from the real system is correctly registered and do not contain too many errors. One of the major limitations is restricted control of variables which may make it hard to discern the real causal relationships, especially since a real network is very complex. This problem can be worked around somewhat if the access to data is good which may allow for diversification, however it is hard to find real operation conditions where only one parameter has changed.

The work presented in report I and paper I aims to describe the whole Swedish railway network, such as infrastructure, timetable, train properties and delays. The rail network is divided into several line sections for which the parameters are calculated. The results are mainly presented as maps showing the network with the parameter values coded in colour or width of the line.

\[ d_e = d_a + d_p + d_s - a_u \]

Several different approaches of analysis exist and involve simulation, optimisation, queue theory and other analytical methods. All methods have their specific strengths and weaknesses and use models with different levels of detail. In general, methods based on less detailed models may be better for drawing general conclusions, which make them suitable tools for long term planning. On the other hand, more detailed models are required to perform thorough studies, but they do also require more data as input and risk generating results that are only valid for a specific setup. This thesis employs both evaluation of empirical data and micro simulation.

4.1 Performance evaluation of the Swedish railway network

Two common reasons for initialising an analysis of a railway network are that either something is not working satisfactory or some prediction needs to be made about the future. If the analysis is about a network currently in use, analysing real operational data may be the first option, rather than using a model. Such an analysis has both advantages and disadvantages. One of the biggest advantages is that if the data comes from a real system and is not the output from a model, it cannot be questioned. Of course, requires that the data from the real system is correctly registered and do not contain too many errors. One of the major limitations is restricted control of variables which may make it hard to discern the real causal relationships, especially since a real network is very complex. This problem can be worked around somewhat if the access to data is good which may allow for diversification, however it is hard to find real operation conditions where only one parameter has changed.

The work presented in report I and paper I aims to describe the whole Swedish railway network, such as infrastructure, timetable, train properties and delays. The rail network is divided into several line sections for which the parameters are calculated. The results are mainly presented as maps showing the network with the parameter values coded in colour or width of the line.
4.1.1 Data preparation

The four sources of data used in the study are:

- **BIS (track information system).** Contains detailed information and location of many of the objects constituting the infrastructure, e.g. signals, switches, speed boards and stations. In this project especially information about the stations is used, such as station coordinates, distance to adjacent stations, station track lengths and simultaneous entry capability.

- **Tidtabellsboken (the Swedish timetable).** Arrival and departure times at stations for all scheduled trains including passenger trains, freight trains, service trains and shunting movements.

- **BANSTAT (train traffic information).** Data including train weight, length and no. of axles is entered into the system by the operator before the train departs from the origin. For each station the train passes, a new entry is generated in the database, repeating the entered values together with the name of each station.

- **TFOR (train delays).** Records delays of scheduled trains. At each passage of a station, the delay is calculated in relation to the scheduled timetable with a resolution of one minute.

Since data from different sources is combined in the calculation of the parameters, it has to be consolidated. In this case, this refers to preparing a common list of stations. Before any calculations can be done, the data needs to be filtered. Some of the systems, like BANSTAT and BIS rely on some data being manually entered which makes errors unavoidable. Even the data from the completely automated system TFOR contains errors. BIS data is manually completed with a few missing links (station-station). Also the data field “simultaneous entry capability” contains a considerable number of errors and is updated according to more reliable data received from train control centres. In the case of BANSTAT, many errors can be removed by calculating the axle load and use knowledge about the maximum permitted axle load and load/m for different lines to identify entries containing to high weights, too few axles or too short lengths.

4.1.2 Calculation of performance indicators

The railway network is divided into line sections that are used in the calculation of the performance indicators. The design of the subdivision is crucial for the usefulness of the indicators. It should neither consist of too short sections, nor too long. Shorter sections give more detailed results, but too many sections may be impractical and make it more difficult to compare line sections in a meaningful way. On the other hand, too long sections make the results too aggregated. The major parameters to consider for defining the sections are traffic patterns and major changes in the standard of the infrastructure, i.e. single track, double track, quadruple track. In the presentation of the results, the Swedish railway network is represented by 123 line sections, figure 4.1.

Since data from different sources is combined in the calculation of the parameters, it has to be consolidated. In this case, this refers to preparing a common list of stations. Before any calculations can be done, the data needs to be filtered. Some of the systems, like BANSTAT and BIS rely on some data being manually entered which makes errors unavoidable. Even the data from the completely automated system TFOR contains errors. BIS data is manually completed with a few missing links (station-station). Also the data field “simultaneous entry capability” contains a considerable number of errors and is updated according to more reliable data received from train control centres. In the case of BANSTAT, many errors can be removed by calculating the axle load and use knowledge about the maximum permitted axle load and load/m for different lines to identify entries containing to high weights, too few axles or too short lengths.

4.1.2 Calculation of performance indicators

The railway network is divided into line sections that are used in the calculation of the performance indicators. The design of the subdivision is crucial for the usefulness of the indicators. It should neither consist of too short sections, nor too long. Shorter sections give more detailed results, but too many sections may be impractical and make it more difficult to compare line sections in a meaningful way. On the other hand, too long sections make the results too aggregated. The major parameters to consider for defining the sections are traffic patterns and major changes in the standard of the infrastructure, i.e. single track, double track, quadruple track. In the presentation of the results, the Swedish railway network is represented by 123 line sections, figure 4.1.
The performance indicators are chosen in accordance to available data and with the intention to describe factors affecting available capacity, used capacity and symptoms of high capacity utilisation. In this case available capacity is represented by infrastructure related indicators to some extent combined with the traffic data, like train length. The timetable is the main input for determining used capacity and delay symptoms of high capacity utilisation, as they are assumed to be capacity dependent. All calculated performance indicators are summarised in the table 4.0 below.
The infrastructure related performance indicators are described by three properties: inter-station distance, no. tracks at stations and their lengths, as well as simultaneous entry capability. Inter-station distance affects the maximum frequency of overtakings and crossings and hence the maximum capacity. On single tracks, simultaneous entry capability reduces the time needed for a crossing. The number of tracks and their lengths can be related to station capacity. On single track, crossing stations with three tracks or more allow simultaneous overtaking and crossing. In addition, redundancy increases for example when faulty freight cars have to be put aside due to e.g. a hotbox. Track lengths on the stations are important for lines used by long freight trains. Too short tracks may increase the effective inter-station distance, hence reducing capacity.

The timetable is used to count the number of trains running on the different line sections during different periods of the selected day. The periods represent the morning and afternoon rush hours together with periods for day and night. The actual time for the peak hour is calculated for each section. By dividing the day into several sections, it is possible to analyse how the traffic is distributed over the day. Running times based on the timetable together with estimated lengths of the line sections makes it possible to classify the sections both according to train speed and to mix of trains of different speeds. Especially the latter has a major impact on capacity consumption and delay propagation.

Data about the length, weight and no. axles can be used to analyse what type of trains are running along different sections. The real lengths of the freight trains combined with the track lengths of the stations indicate if tracks are too short or if it is possible to run longer and therefore fewer trains in order to reduce capacity consumption. The second aspect may be applied to e.g. commuter trains, but then also the platform lengths have to be considered. Calculating gross tons/day gives a hint of the location of the important routes and marshalling yards used by freight trains.

Table 4.0: Performance indicators.

<table>
<thead>
<tr>
<th>Infrastructure</th>
<th>Timetable</th>
<th>Traffic</th>
<th>Delays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single track:</td>
<td>No. Trains per day</td>
<td>Weight (metric tons)</td>
<td>Emissions (g CO2/km)</td>
</tr>
<tr>
<td>Distance between crossing stations (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of trains within 3 km of the crossing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of stations within 3 km of the crossing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-station distance (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single track:</td>
<td>No. Trains per day</td>
<td>Weight (metric tons)</td>
<td>Emissions (g CO2/km)</td>
</tr>
<tr>
<td>Distance between passing stations (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of trains with simultaneous entry capability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The infrastructure related performance indicators are described by three properties: inter-station distance, no. tracks at stations and their lengths, as well as simultaneous entry capability. Inter-station distance affects the maximum frequency of overtakings and crossings and hence the maximum capacity. On single tracks, simultaneous entry capability reduces the time needed for a crossing. The number of tracks and their lengths can be related to station capacity. On single track, crossing stations with three tracks or more allow simultaneous overtaking and crossing. In addition, redundancy increases for example when faulty freight cars have to be put aside due to e.g. a hotbox. Track lengths on the stations are important for lines used by long freight trains. Too short tracks may increase the effective inter-station distance, hence reducing capacity.

The timetable is used to count the number of trains running on the different line sections during different periods of the selected day. The periods represent the morning and afternoon rush hours together with periods for day and night. The actual time for the peak hour is calculated for each section. By dividing the day into several sections, it is possible to analyse how the traffic is distributed over the day. Running times based on the timetable together with estimated lengths of the line sections makes it possible to classify the sections both according to train speed and to mix of trains of different speeds. Especially the latter has a major impact on capacity consumption and delay propagation.

Data about the length, weight and no. axles can be used to analyse what type of trains are running along different sections. The real lengths of the freight trains combined with the track lengths of the stations indicate if tracks are too short or if it is possible to run longer and therefore fewer trains in order to reduce capacity consumption. The second aspect may be applied to e.g. commuter trains, but then also the platform lengths have to be considered. Calculating gross tons/day gives a hint of the location of the important routes and marshalling yards used by freight trains.
The delay data is used to calculate the increase in delay for trains running along the whole section. It is necessary to look at the change in delay rather than absolute delay because most trains travel along several line sections. Based on the assumption that the change in delay is correlated to the length of the line section, the increased delay is normalised by the length of the section to make it comparable between different line sections. Figure 4.2 shows distributions of the delay development on a line section for passenger trains and freight trains. The mean of the distributions are close to zero, and in the case of the passenger trains, slightly smaller than zero which is explained by allowance in the timetable that trains can use to reduce delay. Since the left hand side of the distribution is more closely correlated to available allowance than occurring delays, it is reasonable to focus on the right part, i.e. trains that have increased their delay. The performance indicators based on the delay data is therefore the proportion of the total number of trains with increased delay, and median and standard deviation of the delay increase for the corresponding observations. The median is used rather than the mean to reduce the influence of few observations with very high delays. However, it should be clearly stated that allowances of course also helps to reduce the increase of delays, but by only looking at the positive values the effect of the allowances are somewhat reduced.

4.2 Simulation analysis of double track operation

In general, the process of setting up a simulation model can be divided into four steps: data collection, model implementation, model calibration and model validation. After the model has been calibrated and validated the experiment can be set up, the model applied and the results analysed. In the simulation study presented in this thesis no validation has been done. This is partly because the study involves a fictitious model without corresponding data from real operation, but also because various parts and settings in the simulation tool used, RailSys, has already been validated in earlier projects.
RailSys is an advanced tool for timetable planning and simulation of train operation. The simulation is at a microscopic level with detailed models of infrastructure, timetable, rolling stock and delays. Setting up a simulation in RailSys typically follows the steps below:

1. The model of the infrastructure is created by defining tracks, points, signals, speed boards etc. For many of the objects several properties need to be defined such as gradient and maximum permissible speed (mps) or different interlocking schemes for signals etc. If the objective is to model a real existing railway line, much work lies in obtaining the necessary data and building the detailed model accordingly.

2. Create models of rolling stock that will be used in the simulation. Examples of definable parameters in the rolling stock models are traction force diagram, breaking performance, mass, train length, running resistance etc.

3. Enter the timetable that is to be simulated.

4. Primary delays are defined by entering stochastic distributions. The distributions can be either negative exponential or empirical distributions. Examples of different types of delays that are available are entry delay, dwell time extension, departure delay and running time extension. Preferably, data from real operation can be used to compile the delay distributions. However, it can be tricky to separate primary delays from secondary delays in empirical data that represents total delay.

5. Verify the correctness of the infrastructure and rolling stock models as well as the timetable. If a real timetable and infrastructure is used, much of the verification can be done by checking that running times and allowances are feasible compared to the real timetable. Typically, some test runs also have to be completed in order to check and calibrate the primary delay distributions.

6. Run the simulation with enough number of replications to achieve statistically stable results in the evaluation. The number of replications needed is strongly correlated to the spread of the primary delay distributions applied in the simulation.

7. Evaluate the results.

Normally, the analysis consists of comparing the results from a few simulated scenarios where properties of the infrastructure, timetable or perturbations are varied. Each scenario requires a new simulation where at least some of the steps above have to be completed, which can be very time consuming if the number of scenarios is too high.
To handle experiments with many scenarios, an interface is required to handle input and output from RailSys. In paper II, this is done by transferring data using xml files. It is used to export results from running time calculations in RailSys and certain information about the infrastructure model. The data is then used to create conflict-free simulation-ready timetables as well as perturbation data. The timetable and perturbation data is then imported into RailSys for simulation. This opens up several new possibilities:

- To perform factor analysis of parameters influencing the timetable and perturbations. If more than a few parameters as varied, the number of combinations rises very fast. To setup each scenario by hand would be almost impossible.
- Usually, the simulation results depend to a large degree on the used timetable. This can be reduced by easily generating and simulating several timetables.
- Importing perturbation data allows the applied delays to be controlled in detail. This is not possible if they are created in RailSys, where the delays are generated from a user defined stochastic distribution. An example of an application where this can be used is to study the effects of systematic delays, e.g. a train running with reduced acceleration performance for the entire trip or temporary speed restrictions.
- Creating the timetable yourself outside the simulation tool gives better knowledge about available allowances, which makes it possible to analyse how they are used to recover delay.
- The detailed information about the applied primary delays allows for more detailed analysis of the simulation results and can for example be used for distinguishing secondary delays from primary delays.

In paper II, the interface is used in a factorial experiment with a large number of timetables, three different infrastructure variants, and two levels of primary delays, table 4.1. The infrastructure models consist of one track operated in one direction, thus mimicking the operation of a double with assumed independency of traffic in different directions, with overtaking stations spaced equidistantly. The timetables are defined as cyclic timetables of up to three trains per cycle, table 2. In the scheduling algorithm the timetables are controlled by the starting order of the trains and their headway. The perturbations include three different types of delays, entry delay, running time extension and dwell time extension. All three types are varied coherently for two levels and are based on distributions compiled from empirical data from a real operation, done in an earlier project, Nelldal (2008).

To handle experiments with many scenarios, an interface is required to handle input and output from RailSys. In paper II, this is done by transferring data using xml files. It is used to export results from running time calculations in RailSys and certain information about the infrastructure model. The data is then used to create conflict-free simulation-ready timetables as well as perturbation data. The timetable and perturbation data is then imported into RailSys for simulation. This opens up several new possibilities:

- To perform factor analysis of parameters influencing the timetable and perturbations. If more than a few parameters as varied, the number of combinations rises very fast. To setup each scenario by hand would be almost impossible.
- Usually, the simulation results depend to a large degree on the used timetable. This can be reduced by easily generating and simulating several timetables.
- Importing perturbation data allows the applied delays to be controlled in detail. This is not possible if they are created in RailSys, where the delays are generated from a user defined stochastic distribution. An example of an application where this can be used is to study the effects of systematic delays, e.g. a train running with reduced acceleration performance for the entire trip or temporary speed restrictions.
- Creating the timetable yourself outside the simulation tool gives better knowledge about available allowances, which makes it possible to analyse how they are used to recover delay.
- The detailed information about the applied primary delays allows for more detailed analysis of the simulation results and can for example be used for distinguishing secondary delays from primary delays.

In paper II, the interface is used in a factorial experiment with a large number of timetables, three different infrastructure variants, and two levels of primary delays, table 4.1. The infrastructure models consist of one track operated in one direction, thus mimicking the operation of a double with assumed independency of traffic in different directions, with overtaking stations spaced equidistantly. The timetables are defined as cyclic timetables of up to three trains per cycle, table 2. In the scheduling algorithm the timetables are controlled by the starting order of the trains and their headway. The perturbations include three different types of delays, entry delay, running time extension and dwell time extension. All three types are varied coherently for two levels and are based on distributions compiled from empirical data from a real operation, done in an earlier project, Nelldal (2008).
Table 4.1: Experimental setup.

<table>
<thead>
<tr>
<th>Parameter:</th>
<th>Inter-station distance</th>
<th>Traffic intensity</th>
<th>Perturbation level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[km]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>Low, high</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Cyclic timetables, note that the figures show two cycles. Red: high-speed trains, green: intercity trains, blue: freight trains. Explanation to the values in the table:

Timetable number (train type start order) [heterogeneity, min/100km]

In the experiment, the explanatory variables are the inter-station distance, heterogeneity, number of trains per hour and level of primary delays. The dependent variables are scheduled delay, secondary delay and used allowance. The scheduled delay is a property of the timetable and it is consequently not necessary to perform a simulation to obtain it. The allowance consists of two parts, running time allowance and allowance at stations where trains are scheduled to stop. It is especially the allowance at stations that is dependent on the number of trains and heterogeneity of the timetable, due to the frequency of overtakings.
Secondary delays and how allowances in the timetable are used are dependent of the timetable and applied primary delays. Since the primary delays are modelled by a stochastic process, simulation is needed to obtain them, figure 4.3. The timetable is characterised by no. trains/h and a measure for heterogeneity, i.e. speed mix. The minimum headway referred to in table 4.1 is dependent on type of timetable, i.e. train order, and the inter-station distance. After the timetable is generated, the scheduled delay is evaluated. The available allowance is the sum of running time allowance, allowance at stations and scheduled delay.

### 4.2.1 Heterogeneity, secondary delay and used allowance

In this paper heterogeneity, secondary delay and used allowance are central concepts. All three of them may be defined in several ways. In this project, a measure for heterogeneity is defined that is used to characterise the different types of cyclic timetables and together with no. trains/h used to explain the formation of secondary delays. The difference in free running time of the train types included in the timetable shows good results and is a measure easily applied for the whole line. It does not consider the type of heterogeneity involving unevenly spaced departure/arrival times. However, because the timetables are cyclic with quite few trains/cycle, and the whole line with frequent overtakings is considered, this is of less importance. In addition, the fact that significant entry delays are applied in the simulation reduces the importance of departure related heterogeneity. Since it is based on the free running time, and not the actual scheduled, it is independent of no. trains/h. The difference in running time indicates the difference in speed, and therefore the consequences of faster trains catching up on slower trains on the line sections. Together with the headway, the difference in running time does also hint on the required no. overtakings.

The output from the simulations is arrival and departure delays at the stations and is used to calculate secondary delays and used allowance. In general terms, secondary delays are delays caused by other trains and used allowance is time in the timetable used by trains to reduce delays. However, in practice it becomes more complicated. Consider the example in figure 4.4. The size of the secondary delay can be defined as the time indicated as 1 or 2 in the figure. The difference is whether or not to include the running time allowance that the train could have used if it had not been obstructed. It may not be that important which definition to choose, but it is important that the definition of used allowance is done accordingly to preserve consistency. If the secondary delay is defined as case 1 in the figure, then used running time allowance should be defined as if it is used in the same example.

---

**Figure 4.3: Workflow of the experiment.**

4.2.1 Heterogeneity, secondary delay and used allowance

In this paper heterogeneity, secondary delay and used allowance are central concepts. All three of them may be defined in several ways. In this project, a measure for heterogeneity is defined that is used to characterise the different types of cyclic timetables and together with no. trains/h used to explain the formation of secondary delays. The difference in free running time of the train types included in the timetable shows good results and is a measure easily applied for the whole line. It does not consider the type of heterogeneity involving unevenly spaced departure/arrival times. However, because the timetables are cyclic with quite few trains/cycle, and the whole line with frequent overtakings is considered, this is of less importance. In addition, the fact that significant entry delays are applied in the simulation reduces the importance of departure related heterogeneity. Since it is based on the free running time, and not the actual scheduled, it is independent of no. trains/h. The difference in running time indicates the difference in speed, and therefore the consequences of faster trains catching up on slower trains on the line sections. Together with the headway, the difference in running time does also hint on the required no. overtakings.

The output from the simulations is arrival and departure delays at the stations and is used to calculate secondary delays and used allowance. In general terms, secondary delays are delays caused by other trains and used allowance is time in the timetable used by trains to reduce delays. However, in practice it becomes more complicated. Consider the example in figure 4.4. The size of the secondary delay can be defined as the time indicated as 1 or 2 in the figure. The difference is whether or not to include the running time allowance that the train could have used if it had not been obstructed. It may not be that important which definition to choose, but it is important that the definition of used allowance is done accordingly to preserve consistency. If the secondary delay is defined as case 1 in the figure, then used running time allowance should be defined as if it is used in the same example.
Scheduled stops are modelled by a minimum dwell time and a scheduled dwell time. Perturbations at stops, dwell time extensions, are modelled as a stochastic process. They are added to the minimum dwell time to obtain the minimum time that the train has to stop. If the sum of the minimum dwell time and the dwell time extension is larger than the scheduled dwell time, the train will get delayed (assuming that the train arrives on time). However, if the train arrives late and does not receive a too large dwell time extension, the train can reduce its delay by performing a shorter stop than scheduled.

Exactly how secondary delays and used allowance have been calculated is showed in the equations below.

\[
\begin{align*}
\text{uas}, & \text{ used allowance on station} \\
\text{uasl}_i, & \text{ used allowance on line section} \\
\text{sdl}_i, & \text{ secondary delay on line} \\
\text{ss}_i, & \text{ scheduled stop time} \\
\text{ms}_i, & \text{ minimum stop time}
\end{align*}
\]

\[
\begin{align*}
\text{uasl}_i & = \max(\min(dd_i, al_i), \max(dd_i - da_i, 0), 0) \\
\text{sdl}_i & = \max(da_i, dd_i + uasl_i - re_i) \\
\text{uas}_i & = \min(da_i + de_i, da_i - ms_i) \\
\text{sds}_i & = dd_i + uas_i - da_i - de_i
\end{align*}
\]
5 Results

5.1 Performance evaluation of the Swedish railway network

5.1.1 Infrastructure and traffic

One important factor of the infrastructure that determines the capacity is the distance between overtaking/crossing stations. Especially for single track lines, the inter-station distance is an important factor. The mean value of the inter-station distances on the line sections is shown in figure 5.1 to the right. In general, the distances between crossing stations on single tracks are shorter than those between overtaking stations on double tracks. Also, the possibility to use a side track on the opposite side of a double track for overtakings is limited. Therefore, the practical distances between overtaking possibilities on double tracks are almost twice as long as shown in the figure.

To freight trains, track lengths may be a constraint. The figure to the right shows the mean length of the freight trains. Not only may a short track limit the possibility to run longer and more cost efficient trains, it also influences capacity by increasing the effective inter-station distance if not all crossings or sidings can be used due to short tracks. In the middle figure, the mean track length of each line section has been compared to the lengths of the freight trains passing them. Green means no trains on the evaluation route are longer than the mean track length, and red means that a large percentage of the trains are longer than the mean track length. The range is from 0% to a maximum of 61%. It is evident that some routes have insufficient track lengths. A well-known example is the route between Gallivare and Luleå.
5.1.2 Timetable
The map to the left in figure 5.2 shows the total no. trains per day in black and the same for
only freight trains in green, where the thickness of the lines is proportionate to the number of
trains. It is clear that freight trains dominate in the north. In the south the major flows
of freight trains pass the marshalling yard Hallonberg connecting Gothenburg and Malmö with the
northern parts of the country. Passenger trains dominate around the larger cities Stockholm,
Göteborg and Malmö. The cities are connected by long distance passenger trains via Southern
Main Line (SML) and Western Main Line (WML). The southern parts of SML and WML are
heavily utilised by both passenger trains and freight trains. The average number of trains
operated per day and direction on Swedish single tracks is 11 and on double tracks 83 trains,
but on many sections it is considerably more. The extreme cases are the single track between
Södertälje hamn and Södertälje centrum, which supports 100 trains per day and the double
track between Stockholm södra and Stockholm central with 267 trains. Both sections are very
short, only about 2 km.

5.1.2 Timetable
The map to the left in figure 5.2 shows the total no. trains per day in black and the same for
only freight trains in green, where the thickness of the lines is proportionate to the number of
trains. It is clear that freight trains dominate in the north. In the south the major flows
of freight trains pass the marshalling yard Hallonberg connecting Gothenburg and Malmö with the
northern parts of the country. Passenger trains dominate around the larger cities Stockholm,
Göteborg and Malmö. The cities are connected by long distance passenger trains via Southern
Main Line (SML) and Western Main Line (WML). The southern parts of SML and WML are
heavily utilised by both passenger trains and freight trains. The average number of trains
operated per day and direction on Swedish single tracks is 11 and on double tracks 83 trains,
but on many sections it is considerably more. The extreme cases are the single track between
Södertälje hamn and Södertälje centrum, which supports 100 trains per day and the double
track between Stockholm södra and Stockholm central with 267 trains. Both sections are very
short, only about 2 km.
5.1.3 Delays
The left part of figure 5.3 shows the proportion of passenger trains that have increased their delays. The middle and right figures show the median of the delay increase per 100 km for passenger and freight trains respectively. The proportion of freight trains is within the same interval as for the passenger trains, but the median is much higher for the freight trains. The reason for this is the larger spread of the distribution for freight trains, figure 4.2. Line sections with a very high proportion of delayed trains indicate a systematic problem that affects all passing trains. This can be sections where the timetable has low allowances or congested areas where secondary delays easily occur and it is hard to recover due to congestion. Especially on congested single track lines, once a train is delayed, additional delays are probable due to frequent crossings. Some of the line sections with the highest proportion of delayed trains, up to 90%, are due to temporary speed restrictions being active during the whole period of measurement. In these cases, the timetable have not been adjusted to accommodate for longer running times, i.e. the running time allowance on these sections are non-existent or even negative.

There are some line sections where the proportion of delayed trains is small, but the median is high, thus indicating that relatively few trains get large delays. A possible explanation is that it is a specific type of trains that get delayed, like long distance passenger trains sharing the same tracks with slower local trains. The long distance trains are then a smaller proportion of the total no. trains, compared to sections without local services, but may get delayed by the slower local trains. This effect can be observed on Mälar line close to Stockholm and on Western Main Line close to Goteborg and Stockholm. Approaching the cities, the proportion of delayed trains are at first high and the median delay low, but closer to the cities on the lines where local services operate, the situation is the opposite.

There are some line sections where the proportion of delayed trains is small, but the median is high, thus indicating that relatively few trains get large delays. A possible explanation is that it is a specific type of trains that get delayed, like long distance passenger trains sharing the same tracks with slower local trains. The long distance trains are then a smaller proportion of the total no. trains, compared to sections without local services, but may get delayed by the slower local trains. This effect can be observed on Mälar line close to Stockholm and on Western Main Line close to Goteborg and Stockholm. Approaching the cities, the proportion of delayed trains are at first high and the median delay low, but closer to the cities on the lines where local services operate, the situation is the opposite.
The performance indicators are defined with the intention to describe properties associated with capacity, where the infrastructure indicators indicate available capacity and timetable and traffic indicators used capacity. Based on the assumption that delays are correlated to capacity utilisation, a correlation analysis is performed where the indicators of available and used capacity are used to explain the delays. The correlation analysis is separated for single/double tracks, passenger/freight trains and for the proportion of delayed trains and their used capacity are used to explain the delays. The correlation analysis is separated for single/double tracks, passenger/freight trains and for the proportion of delayed trains and their median delay. No significance is found for the double track sections, probably due to too few line sections, while the results for the single tracks are summarised in the table 5.1.

5.1.4 Correlation analysis

Figure 5.3: Left, proportion passenger trains with increased delay. Middle, median of the increased delay for passenger trains. Right, median of the increased delay for freight trains.

The performance indicators are defined with the intention to describe properties associated with capacity, where the infrastructure indicators indicate available capacity and timetable and traffic indicators used capacity. Based on the assumption that delays are correlated to capacity utilisation, a correlation analysis is performed where the indicators of available and used capacity are used to explain the delays. The correlation analysis is separated for single/double tracks, passenger/freight trains and for the proportion of delayed trains and their median delay. No significance is found for the double track sections, probably due to too few line sections, while the results for the single tracks are summarised in the table 5.1.
Table 5.1: Parameter analysis (single tracks), significant at the 0.05 level.

<table>
<thead>
<tr>
<th>Train type</th>
<th>Delay type</th>
<th>Parameter</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passenger</td>
<td>Proportion</td>
<td>Total nr of trains/day</td>
<td>0.44</td>
</tr>
<tr>
<td>Passenger</td>
<td>Proportion</td>
<td>Mean speed (all trains)</td>
<td>0.42</td>
</tr>
<tr>
<td>Passenger</td>
<td>Proportion</td>
<td>Speed mix (all trains)</td>
<td>0.29</td>
</tr>
<tr>
<td>Passenger</td>
<td>Median</td>
<td>Route length</td>
<td>-0.58</td>
</tr>
<tr>
<td>Passenger</td>
<td>Median</td>
<td>Share of passenger trains with &gt; 12 axles</td>
<td>0.24</td>
</tr>
<tr>
<td>Passenger</td>
<td>Median</td>
<td>Std of the inter-station distance</td>
<td>-0.41</td>
</tr>
<tr>
<td>Freight</td>
<td>Proportion</td>
<td>Total nr of trains/day</td>
<td>0.57</td>
</tr>
<tr>
<td>Freight</td>
<td>Proportion</td>
<td>Share of stations with at least 3 tracks</td>
<td>-0.46</td>
</tr>
<tr>
<td>Freight</td>
<td>Median</td>
<td>Route length</td>
<td>-0.37</td>
</tr>
<tr>
<td>Freight</td>
<td>Median</td>
<td>Freight train mass</td>
<td>0.37</td>
</tr>
<tr>
<td>Freight</td>
<td>Median</td>
<td>Max inter-station distance</td>
<td>-0.37</td>
</tr>
</tbody>
</table>

Some of the results are difficult to interpret but point to areas where the indicators and the analysis can be improved. Especially the results for the median delays are questionable and show that the route length have a negative correlation to the median delay of both passenger and freight trains, despite the fact that the mean delay is normalized by the route length to eliminate this factor. The explanation is the low resolution (1 minute) of the empirical delay data. It follows that the minimum accumulated delay for a train is 1 minute (given that the train has received an additional delay) and that the median accumulated delay for all trains must be at least 1 minute. This is a problem for short line sections where the median delay hits the 1 minute limit, with the consequence that the increased delay/km will be very high for the shortest evaluation routes, hence the negative correlation. The problem is illustrated in figure 5.4.

Figure 5.4: The solid line shows the effect of the route length of the minimum possible delay increase.

There are several reasons beside the one mentioned above why the correlation analysis is hard to perform. Maybe the biggest is the limited no. observations of the aggregated indicators where several indicators covariate.
5.2 Simulation analysis of double track operation

The bar graph below summarizes the results of one of the timetables consisting of freight trains, IC trains and high speed trains, i.e. one of the more heterogeneous timetables. The graph shows clearly how both the timetable and the trains in operation are affected when traffic density is increased. The figures are mean values for all train types combined. The bars showing the available allowance include scheduled delay, hence the dramatic increase in available allowance at stations as traffic density grows and overtakeings become more frequent. The secondary delays at stations increase somewhat for every increment in traffic density while the secondary delays on line sections increase slowly at first and then more dramatically at the highest two levels. Secondary delays at stations are mainly caused by low priority trains waiting to be overtaken by high priority trains, while on line sections, trains tend to interfere with other trains more freely, regardless of priority.

It is also evident in the figure that the allowance at stations that is used to reduce delay increases with higher traffic densities, while the used running time allowance remains approximately constant. The main reason for this is the increase in available allowance at stations. For the first four timetables, the increase in used allowance manages to compensate for the increase in secondary delay, and it is not until the final two timetables that the exit delay starts to increase. All in all, the graph shows how allowance and delays interact and the result thereof, i.e. exit delay.

One methodological aspect apparent from figure 5.5 is that all types of primary delays are as good as constant for all simulations. This is intended and shows that enough replications have been simulated to achieve stable mean values. Another is that the bars showing the delays and used allowance sums up to the same value, which shows that the definitions of secondary delays are consistent with the definitions of used allowance.

5.2.1 Heterogeneity

One of the main objectives of paper II, is to analyse how heterogeneity affects delays and capacity. The heterogeneity measure described in section 4.2.1 together with number of trains per hour are used to explain secondary delays, available allowance and used allowance for a given infrastructure variant and level of primary delay. In the scheduling scheme used in this paper, faster trains are given absolute priority over slower, hence they never receive any scheduled delay due to slower trains. This means that the heterogeneity does not only
influence the scheduled delay applied to the slower trains and delay propagation, but it does also severely reduce the number of trains it is possible to schedule. Figure 5.6 to the left shows some results for one timetable classified as completely homogenous, just one train type, and one as heterogeneous (same as in figure 5.5). The difference in behaviour is clear. For the heterogeneous timetable, the secondary delays increase but are at first compensated by higher use of allowance, made possible by the rapid increase of available allowance, before also the exit delay starts to go up. For the homogenous timetable, the increase of available allowance is already used at the beginning and no extra allowance is given at higher traffic densities. The reason why so much of the allowance is used already at low traffic densities is that most of the allowance at stations is used to compensate for applied dwell time extensions. At the high perturbation level almost 40% of the trains receive a longer dwell time extension than the available allowance. The result is that there is only room for a marginal increase of used allowance and consequently the development of the exit delay follows that of the secondary delay quite well. However, the increase in secondary delay and exit delay is not as large as in the heterogeneous case, despite much higher traffic densities.

![Figure 5.6: Left: Timetable type 1 (homogenous, solid line) and 10 (heterogenous, dashed line), 20 km interstation-distance and high perturbation level. Right: All timetables, 20 km interstation-distance and high perturbation level. Grey contours indicate the 95% prediction intervals of the surface fit. The right part of figure 5.6 shows secondary delay as a function of no. trains/h and heterogeneity. The delay is shown as contours of a surface fitted to the simulated results of all 84 timetables simulated for the current infrastructure variant and level of primary delay. The grey dashed and dash/dotted contours shows the 95 % prediction intervals. The data showing the secondary delays in the left figure is a subset of the data in the right figure. The homogeneous and heterogeneous timetables in the left figure have heterogeneities 0 and 18.5 min/100km. Comparing the two figures shows that the surface fits the data of the homogenous timetable rather good. For the heterogeneous timetable, the fit is worse close to maximum capacity. This is partly explained by that there are many timetables in the experiment with heterogeneity 18.5 min/100km or close, table 4.2, with different sensitivity to delays. The red markers in the right diagram show the location of the simulated timetables and it is obvious that the spread of the timetables are not optimal. This is due to the limitation of using cyclic timetables of only up to three trains per cycle and three train types. Even if the experiment was not designed with the right diagram in mind, it is the reason why a
The impact of primary delays on the creation of secondary delays is shown in figure 5.7, and corresponds to the difference of the two contour plots. For a given amount of secondary delay, going from a low to a high level of primary delay corresponds to a quite severe reduction in the number of trains it is possible to run. In terms of no. lost trains/h, timetables close to maximum capacity are more sensitive to primary delays, which is natural due to smaller buffer times. Homogeneous timetables at high traffic densities are sensitive to primary delays due small buffer times. Heterogeneous timetables have dependencies between trains caused by overtakings that will transfer delays.

In the experiment all types of primary delays are varied together for two levels, high and low. Consequently, it is not possible to determine the effect of the different types of primary delay.

5.2.2 Primary delay

The impact of primary delays on the creation of secondary delays is shown in figure 5.7, and corresponds to the difference of the two contour plots. For a given amount of secondary delay, going from a low to a high level of primary delay corresponds to a quite severe reduction in the number of trains it is possible to run. In terms of no. lost trains/h, timetables close to maximum capacity are more sensitive to primary delays, which is natural due to smaller buffer times. Homogeneous timetables at high traffic densities are sensitive to primary delays due small buffer times. Heterogeneous timetables have dependencies between trains caused by overtakings that will transfer delays.

In the experiment all types of primary delays are varied together for two levels, high and low. Consequently, it is not possible to determine the effect of the different types of primary delay.

5.2.3 Inter-station distance

Looking at the sum of secondary delays for all train types, the inter-station distance has practically no effect, left part of figure 5.8. However, if the results are analysed separately for different train types, it becomes apparent that faster trains gain from shorter inter-station distances, while slower lose, see example to the right in figure 5.8.

The main reason that slower trains gain from longer inter-station distances is that the number of overtakings required remains approximately the same, or will even increase slightly. Since the 40 km infrastructure variant has fewer stations, a larger proportion of the stations will have scheduled overtakings. The consequence is lower secondary delays at stations for slow trains in the 40 km case, due to fewer overtakings taking place at stations without scheduled stops. Faster trains suffer from being caught behind slower trains for a longer distance in the 40 km case, compared to the 20 km case. Also because the stations are fewer in the 40 km case, the total capacity for overtakings is lower, which will affect the faster trains.

The main reason that slower trains gain from longer inter-station distances is that the number of overtakings required remains approximately the same, or will even increase slightly. Since the 40 km infrastructure variant has fewer stations, a larger proportion of the stations will have scheduled overtakings. The consequence is lower secondary delays at stations for slow trains in the 40 km case, due to fewer overtakings taking place at stations without scheduled stops. Faster trains suffer from being caught behind slower trains for a longer distance in the 40 km case, compared to the 20 km case. Also because the stations are fewer in the 40 km case, the total capacity for overtakings is lower, which will affect the faster trains.
5.2.4 Train types

Several aspects have to be considered when capacity is defined and conditions have to be applied to both the timetable and to the train operation. For the timetables, these conditions are derived from demand and may for example include clock-face timetables, and scheduled delay. Level of acceptance for delays might be the most important condition on the train operation. In our case, relevant limits have to be set for scheduled delays and operational delays in order to determine the maximum capacity under different conditions. As has been shown earlier, both the scheduled and operational delay goes up when traffic density is increased. In this case the slower train types have lower priority both in the scheduling procedure and in the simulation. The consequence is that slower trains receive both scheduled delay and operational delay, while faster trains suffer only from operational delay. This indicates that the results have to be separated according to train type in the analysis.

Worth commenting is also the fact that the scheduled delay for freight trains decrease when traffic density increase from 7.4 to 8.8 trains/h. This is an effect of using cyclic timetables and an infrastructure with the overtakings stations spaced equidistantly. A shorter headway may cause a better timing at the overtakings, i.e. the freight trains have to wait for a shorter time before the high-speed trains arrive, while the number of overtakings required remains the same. This is supported by the fact that the running time allowance, which is 6% of the scheduled running time, remains exactly the same for the two timetables.

Available allowance. As mentioned before, it is only the freight trains that receive scheduled delay due more frequent overtakings at higher traffic densities. The scheduled delay is substantial and becomes as much as 39 min/100km, outside the figure, which corresponds to an increase in scheduled running time of 65%. The scheduled delay is closely connected to the scheduling scheme. It is possible that if small scheduled delays are accepted also for the high priority trains, the scheduled delay for the lower priority trains would decrease significantly.

Worth commenting is also the fact that the scheduled delay for freight trains decrease when traffic density increase from 7.4 to 8.8 trains/h. This is an effect of using cyclic timetables and an infrastructure with the overtakings stations spaced equidistantly. A shorter headway may cause a better timing at the overtakings, i.e. the freight trains have to wait for a shorter time before the high-speed trains arrive, while the number of overtakings required remains the same. This is supported by the fact that the running time allowance, which is 6% of the scheduled running time, remains exactly the same for the two timetables.

Figure 5.8: Left, effect of inter-station distance on the total secondary delay.
Right, effect of inter-station distance for one type of timetable, separated for the different train types.
Used allowance. The large difference in used allowance between the train types is explained by the difference in available allowance. For freight trains the increase in used allowance at stations is more than 5.5 minutes, which more than well covers for the increase in secondary delays. For the high-speed trains there is almost no increase in used allowance at stations, and a very small increase of used running time allowance. The limited increase is explained by that most of the allowance is already used, even at low traffic densities. Compare with the homogenous timetable in figure 5.6.

Secondary delays. Freight trains receive most of their secondary delay at stations while the high-speed trains get it on line section. The reason for this is that lower priority trains have to wait at the stations to be overtaken by faster high priority trains. The efficient dispatching and the fact that dwell time extensions rather than departures delays have been used in the simulation has the effect that the high priority trains get next to no secondary delays at stations. Looking at the secondary delays in total, freight trains get some delays even at low traffic densities. It increases with traffic density, but seems to level out somewhat. For the high-speed trains, the secondary delays are at first almost non-existent, but at around 5 trains/h the secondary line delays start to increase quite fast. The rapid increase is probably explained by the limited capacity of the two track stations that only allow one train to be overtaking at a time. At 7.4 and 8.5 trains/h, overtakings are scheduled at every station. Exit delay. The exit delay does also differ between the two train types and is explained by the difference in used allowance. The development of the exit delay of the high-speed trains follow quite well that of the secondary delays, which is natural since nothing else changes much. The freight trains however, manage to keep the exit delay constant, or even reduce it slightly, as the traffic density increases. Even though figure 5.9 does not show the same timetable as figure 5.5, the behaviour of the involved train types are the same. In figure 5.5, the exit delays remain stable at first, and then start to increase in union with the secondary delay on line sections. Looking at figure 5.9, this behaviour can now be explained by that it is the secondary line delays of the high priority trains that starts to go up, and since they cannot use any allowance for recovery, so does the exit delay.

Exit delay. The exit delay does also differ between the two train types and is explained by the difference in used allowance. The development of the exit delay of the high-speed trains follow quite well that of the secondary delays, which is natural since nothing else changes much. The freight trains however, manage to keep the exit delay constant, or even reduce it slightly, as the traffic density increases. Even though figure 5.9 does not show the same timetable as figure 5.5, the behaviour of the involved train types are the same. In figure 5.5, the exit delays remain stable at first, and then start to increase in union with the secondary delay on line sections. Looking at figure 5.9, this behaviour can now be explained by that it is the secondary line delays of the high priority trains that starts to go up, and since they cannot use any allowance for recovery, so does the exit delay.
6 Conclusions
This thesis discusses results from two studies based on different methodologies. The first is an analysis of recorded data from real operation on a real network and the second employs simulation of fictitious timetables on fictitious double track railway lines. Both methods have their strengths and weaknesses. Analysing data from an entire network can point to problematic sections with high capacity consumption and hint on possible actions to improve the situation. However, due to the complexity of the system where the timetable and layout of stations are examples of factors of great influence, the data need to be studied on a highly detailed level if causal relationships are to be established. With highly detailed level means individual trains and their timetables, as well as track layout at stations, etc.

In the attempt to define maximum capacity, it is clear that scheduled delay have to be taken into consideration. It is dependent on capacity utilisation and since it affects the travel time, it is undesirable if it becomes too high. However, its positive effects on delay reduction should also be considered in the evaluation. By looking at the development of the actual delay as trains run from origin to destination, the positive effect of scheduled delay is included as well as the negative of secondary delays. The delay development can be calculated as the difference in entry delay and exit delay of the trains. However, this method is very sensitive to if an overtaking is scheduled right before the destination, which would reduce the exit delay significantly. Thus it is more sensible to look at the trend of the delay using observations at all intermediate stations as well. For this reason a robust fit is a used to find the trend of the delay, i.e. the delay development. For each timetable, the delay development is calculated separately for all train types.

The delay development together with the scheduled delay can now be used to define when a timetable has reached its maximum capacity. Typically it is the delay development for the high priority trains or the scheduled delay of the low priority trains that set the capacity limit. Values for acceptable delay development and scheduled delay may be different for different train types, where the greatest difference is between passenger trains and freight trains. The intention with figure 6.10 is to show how capacity depends on timetable heterogeneity, accepted scheduled delay and delay development. It shows the delay development in colour and three lines corresponding to different levels of accepted scheduled delay. In case the timetable consists of several train types, the value for the delay development is the maximum of the included train types. The solid line indicates the upper limit of what is possible to schedule under the given circumstances, and still maintain a timetable free of conflicts. Hence values to the right of this line are not valid. No conditions are set on the scheduled delay for the solid line. The dashed line indicates the limit if scheduled delays of up to 40, 20 and 5% are accepted for freight, intercity and high-speed trains respectively. For the dash-dotted line the corresponding values are 30, 10 and 5%. The axes show traffic density and heterogeneity.
Figure 6.10 shows six different scenarios that correspond to the combinations of high and low perturbation levels and all infrastructure variants. Together they summarise the effect on capacity of all factors from the double track simulation experiment. Several interesting observations can be made:

- It is possible to schedule more trains/h with shorter inter-station distances. This effect increases with heterogeneity and is natural since the minimum headway for two trains of different speeds is directly dependent on the inter-station distance. However, the extra train slots come at a high price of scheduled delay, which is seen as the difference between the solid line and the other lines.

- The impact of the inter-station distance on the limits for scheduled delay is small, i.e. the dashed and dashed-dot line.

- For the low primary delays the capacity is limited by the acceptance of scheduled delay. For high level of primary delays, also delay development becomes a limiting factor.

- Inter-station distance has some effect on delay development. The difference between the 20 km and 40 km variants is around 2 trains/h if the primary delay level is high.

- For heterogeneous timetables, the transition from a stable (green) to a highly unstable timetable (red) goes quite fast. It is probably at this point that the secondary line delays for the high priority trains starts to grow, as previously discussed.

Paper II shows the importance of understanding how allowances affect delays. For low priority trains receiving increased scheduled delay as capacity utilisation increases, actual delays do not increase even though secondary delays are higher. In reality both high-speed trains and slower trains may receive scheduled delay in congested situations, which would help reduce the delays also for high-speed trains, but then at the cost of increased travel time. Consequently, in order to get the whole picture, allowances and scheduled delay needs to be included in the analysis of operational delays, also when data from a real network is analysed.
Figure 6.10: Left column, low perturbation level. Right column, high perturbation level. Each row corresponds to one infrastructure variant. The color indicates the delay development. Note that the color bar saturates at -0.5 and 2 min/100km, and that lower and higher values are possible. Accepted scheduled delay for the dashed and dash-dotted lines are 40, 20, 5% and 30, 10, 5% of free running time for freight, intercity and high-speed trains respectively.
7 Contribution of the thesis

In paper I data from several databases containing information about the Swedish rail network infrastructure, timetable, traffic and delays are combined to calculate several performance indicators. Even though it proved hard to correlate total delays to capacity utilisation, several of the key performance indicators proved useful in highlighting problematic conditions like traffic speed mix and insufficient track lengths at crossing and overtaking stations.

In paper II a method is developed that expands the potential of using the micro simulation tool RailSys to perform more general and less timetable dependent capacity analysis. The planned timetable and perturbations have a decisive impact on the performance of the train operation. This implies not only that both the timetable and the perturbations need to be modelled in detail, but also the need for a dispatching functionality. These are all areas where micro simulation excels.

However, it also follows that it may be hard to draw general conclusions about e.g. the capacity of a railway line by just analysing one or a very few timetables and levels of perturbations. For this reason a program is constructed in paper II that allows timetables and perturbation data to be imported into RailSys. The interface, together with timetable and perturbation generators, makes it possible to simulate hundreds of scenarios, which would be practically impossible to do manually. This way a more general analysis of a given railway line can be performed, without losing the advantage of the detailed analysis that micro simulation can offer.

The second aspect of contribution in paper II is the analysis of the simulation results. The analysis is focused on explaining the behaviour of double track train operation under different conditions. The formation of secondary delays and the how much of the available allowance in the timetable is used to recover delays are evaluated. Both of these are difficult to estimate from delay data without detailed knowledge about primary delays and available allowance. In paper II this information is available as it is part of the input to the simulation tool, and thus secondary delays and used allowance can be estimated on both line sections and at stations with high accuracy. This data together with the primary delays and available allowance gives a comprehensive understanding of how different types of timetables react to increased capacity utilisation and explain the outcome in term of exit delay.

Finding general and applicable definitions for maximum capacity and heterogeneity remains a difficult task. The heterogeneity measure introduced in paper II, difference in free running time, shows a high explanatory value for the creation of secondary delays. It relates only to the difference in speed, not unevenly distributed departure times. However, the results suggest that for analysis of whole railway lines with cyclic timetables with short cycles and significant entry delays, difference in speed is the important factor. Also, the simulation results are used to demonstrate a way to define maximum capacity, which considers inter-station distance, primary delays, heterogeneity as well as level of acceptable scheduled and operational delays.
8 Future research

The analysis performed in paper I can be improved with respect to allowances. Firstly, the timetable can be used to estimate the available allowance at stations. Available running time allowance can be estimated by comparing the scheduled running time with the shortest of the actually realised running times estimated from empirical delay data. This has to be done for each individual train in the timetable and requires delay data for all stations and for a long time period. Also another possibility is to complement the real data with running time calculations performed in e.g. RailSys. This, however, requires detailed data about what train types are used to operate different trains in the timetable. The information can be used to improve the correlation analysis performed in paper I. Also, the delay data can be separated into several sets, in order to cover different operating conditions. For example, the data can be split according to time of day or day of week, to diversify with respect to capacity utilisation, figure 8.1. However, one should be aware that other conditions might also change, like traffic mix and primary delays due to e.g. maintenance works during nights and weekends.

Also another possibility is to complement the real data with running time calculations performed in e.g. RailSys. This, however, requires detailed data about what train types are used to operate different trains in the timetable. The information can be used to improve the correlation analysis performed in paper I. Also, the delay data can be separated into several sets, in order to cover different operating conditions. For example, the data can be split according to time of day or day of week, to diversify with respect to capacity utilisation, figure 8.1. However, one should be aware that other conditions might also change, like traffic mix and primary delays due to e.g. maintenance works during nights and weekends.

Figure 8.1: Traffic work for the Swedish railway network for different days of the week. Source: Planned timetable and BIS, time period: 2008.10.06-12.

Leaving the delays behind, the other performance indicators can be used to estimate capacity utilisation. From paper II, the most significant term in explaining secondary delays and delay development is no. trains multiplied by the heterogeneity. With this in mind, a quick estimation of used capacity for the Swedish network can be calculated by multiplying the no. trains/day and a heterogeneity measure (95 percentile/10 percentile). To make the results comparable between single and double tracks, the product above is divided by 4 for all double track sections. The reason is that in the most general approach, the capacity of a double track is four times that of a single track. The result is shown in figure 8.2 below together with results from the yearly analysis of capacity utilisation and limitations conducted at the Swedish Transport Administration (Trafikverket). The results are similar, however the mathematical model employed by Trafikverket is more complex and takes more factors into consideration, like for example inter-station distance and demand Grimm (2008). A possibility of future research is to use the key performance measures calculated in paper I to calculate the capacity consumption, like the example below, and calibrate it against the results from Trafikverket’s calculations. Also, there might be need to separate long distance services from local services in the analysis of the delays. Typically long distance services may be more sensitive to delays because they are the faster trains close to the cities where local trains operate, and they travel longer and may consequently accumulate delays along the way.
In paper II the possibility of importing timetables and perturbations directly into RailSys opens up several new opportunities. For example, can the impact of primary delays be studied in detail. In paper II all types of primary delays are varied together, which makes it impossible to distinguish their individual effects. A first step would be to vary the different types of primary delays individually. Another possibility is to apply systematic primary delays to mimic specific delay causes, like temporary speed restrictions or trains running with reduced performance. A third option is to gradually increase a specific running time extension or dwell time extension for a specific train. The effect can then be seen as changes in secondary delays up and down stream. This gives insight into how delays spread in the network.

In the algorithm programmed in paper II to generate the timetables, trains have to be given unique priorities. The trains are then entered into the timetable according to these priorities, and a high priority train is never given any scheduled delay for the benefit of a low priority train. This gives timetables where faster high priority trains are more sensitive to delays and low priority train get more scheduled delay than may be realistic. A timetable generator that minimises, at least locally, the scheduled delay for all train types would give more flexible and realistic timetables.

Another option is to let RailSys create the timetables. This can be done by using entry delays to systematically manipulate the departure times of the trains by, and then let the dispatching in RailSys create the timetables for you. Each simulated cycle represents a new timetable. The created timetables can then be evaluated according to some criteria, and the best alternatives selected to be analysed further by simulation with stochastic primary delays. This allows for more complex infrastructures and traffic patterns to be analysed, which makes it easier to apply the method on real railway lines. This can be combined with the possibility of
utilising RailSys to calculate the capacity utilisation according to UIC 406 (2004) method, and analyse how it correlates to simulation results.

In order to make the results from the simulation analysis more useful in for example investment planning, socioeconomic costs can be assigned to scheduled delays and operational delays. These costs can be differentiated according to train types and weighted differently depending on the delay size. Costs for infrastructure investments can then be compared to the benefit in form of reduction of scheduled and operational delay.
9 References


