ROYAL INSTITUTE
OF TECHNOLOGY

Railway capacity analysis

-

Methods for simulation and evaluation of timetables,
delays and infrastructure

Anders Lindfeldt

Doctoral Thesis in Infrastructure

Stockholm 2015

TRITA-TSC-PHD 15-002
ISBN 978-91-87353-65-9

KTH Royal Institute of Technology
School of Architecture and the Built Environment
Department of Transport Science

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 transportation. To meet this new demand, railways play an important role. Today, the capacity of the Swedish rail network is not upgraded at the pace 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 maximum capacity is reached when the marginal gain of operating one extra train is lower than the costs in terms of longer travel times and increased sensitivity to delays.

Several different 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 train delays. 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, traffic heterogeneity, primary delays and inter-station distance on secondary delays, used timetable allowance and capacity. The analysis gives an in-depth understanding of the mechanisms of railway operation on double-track lines.

A simulation model for strategic capacity evaluation, TigerSim, is developed that can be used to speed up and improve capacity planning and evaluation of future infrastructure and timetables designs on double-track railway lines. For a given infrastructure and plan of operation, the model can be used to generate and simulate a larger number of timetables. This gives two major advantages:

- Using many timetables makes results general
- It is possible to consider both static and dynamic properties of the timetables in the capacity analysis.

The first aspect is especially useful in the evaluation of future scenarios as the timetable then often is unknown. The second is an advantage since an improvement in capacity can be measured in a combination of increased frequency of service, shorter travel time and reduced delays. The output of the model can either be used to directly determine capacity from a quality of service perspective, or used as input to cost-benefit analysis (CBA).

Acknowledgements

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

I would like to thank my supervisors Bo-Lennart Nelldal, Lars-Göran Mattsson and Markus Bohlin 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 and Pär Johansson at Trafikverket, members of the reference group, and the “railway simulation team” at Trafikverket: Per Köhler, Magnus Backman, Johan Mattisson. Thanks to Anders Peterson and Mats Berg for valuable comments regarding the thesis. Thanks to Hans Sipilä, my roommate during all these years, for the endless discussions about work and everything. It is possible that without you, I would have finished my PhD faster, but more probably not at all. Thanks to Jennifer Warg, Behzad Kordnejad, Jiali Fu, Oskar Fröidh, Anders Lindahl, Hans E Boysen, Josef Andersson and Fredrik Hagelin for all the stimulating discussions and their valuable inputs. Thanks to all colleagues and fellow PhD students at Department of Transport Science for making work fun. Special thanks to Olov Lindfeldt and Leif Broberg for introducing me to the field of railway capacity.

To my family: mother, father, brothers, uncles, aunt and friends, without you it would not have been possible. Min kone Nadire, köp rachmet jinim! Maen sinni jikk soymaen xikirim apigim!

Stockholm, August 2015

Anders Lindfeldt

List of Publications

Papers

I. Lindfeldt, A., 2010. A study of the performance and utilization of the Swedish railway network. Published in: Proceedings of the First International Conference on Road and Rail Infrastructure, Opatija, Croatia.

II. Lindfeldt, A., 2011. Investigating the impact of timetable properties on delay propagation on a double-track line using extensive simulation. Published in: Proceedings of Railway Engineering, 11th International Conference, London, UK.

III. Lindfeldt, A., Sipilä, H., 2014. Simulation of freight train operations with departures ahead of schedule. Published in: Proceedings of the 14th International Conference on Railway Engineering Design and Optimisation. (CompRail XIV), Rome, Italy.

IV. Lindfeldt, A., 2013. Heterogeneity Measures and Secondary Delays on a Simulated Double-Track. Published in: Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen2013), Copenhagen, Denmark.

V. Lindfeldt, A., 2015. Validation of a simulation model for capacity evaluation of double-track railway lines. Published in: Proceedings of the 6th International Seminar on Railway Operations Modelling and Analysis (RailTokyo2015), Tokyo, Japan.

VI. Lindfeldt, A., 2015. Scheduled waiting time and delay in capacity evaluation of double-track railway lines. Submitted to Journal of Rail Transport Planning & Management

Related publications not included in thesis

- Nelldal, B-L., Lindfeldt, A., Lindfeldt, O., 2009. Kapacitetsanalys av järnvägsnätet i Sverige, delrapport 1. Capacity analysis of the Swedish rail network, part 1. In Swedish.

- Lindfeldt, A., 2009. Kapacitetsanalys av järnvägsnätet i Sverige, delrapport 2. Capacity analysis of the Swedish rail network, part 2. In Swedish.

- Lindfeldt, A., 2014. Kapacitetsutnyttjande i det svenska järnvägsnätet. Uppdatering och analys av utvecklingen 2008-2012. Capacity utilisation of the Swedish rail network. Update and analysis of the development 2008-2012.

Contents

1 Introduction

Railway is an attractive mode of transportation with a wide range of applications. It is used to provide high capacity local transportation of passengers in large cities, haul large quantities of ore from inland located mines to seaside ports and to connect cities by comfortable high-speed services offering short travel times from city centre to city centre, just to mention a few. It is energy efficient and it can be powered by renewable energy sources, hence its attractiveness is bound to increase further as awareness increase of the issues associated with air pollution and climate change.

The demand for railway transportation is steadily increasing around the world (UNECE, 2014). The increase in demand generates increase in traffic load. Many railway lines are already used close to their maximum capacity and in order to meet the new demand, actions need to be taken. Such actions include building new railway infrastructure, upgrade existing infrastructure or use existing infrastructure more efficiently. Constructing new railway infrastructure is expensive, and it is therefore of importance that the right actions are taken at the right time. This in turn requires an understanding about how the railway system works and responds to increased capacity utilisation.

Analysing and describing railway capacity is a multifaceted task. It involves several complex systems, e.g. railway infrastructure, rolling stock, timetable and human behaviour. Some of the factors influencing capacity are: number of tracks connecting the stations, station track layout, signalling system, train performance, speed difference between train services, market demand, reliability and delay acceptance of railway customers. Because if its complexity, railway capacity can be defined in many different ways. Barter (2008), quoting (Nock, 1980), gives a good general definition of railway capacity that includes the most important aspects:

> *The number of trains that can be incorporated into a timetable that is conflict free, commercially attractive, compliant with regulatory requirements, and can be operated in the face of anticipated levels of primary delays whilst meeting agreed performance targets.*

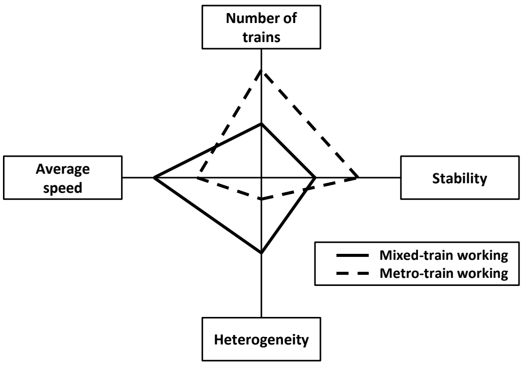


Figure 1.1: Capacity balance (UIC, 2004).

There is no simple way to tell what the capacity of a railway infrastructure is because it depends to such a high degree on how it is used (UIC, 2004). Figure 1.1 shows that capacity is a balance between number of trains, average speed, stability and heterogeneity. Capacity is the length of the chord connecting the four axes. It shows that railway capacity is a trade-off between quantity and quality, i.e. between number of trains that are operated and how much delays they will experience. Increased traffic load leads to higher sensitivity to delays with more secondary delays propagating from train to train. It also makes it harder to construct timetables with attractive train paths, especially if it is a single track line or a double-track line with heterogeneous traffic mix. Scheduled waiting time (SWT) increases when more train slots are scheduled and crossings and overtakings become more frequent. There is a conflict of interest between adding new train paths to meet higher demand and maintaining the quality of the already scheduled trains. The need to solve this conflict accurately grows when the market for railway operation is deregulated and service operators have to be denied train slots due to capacity constraints.

This doctoral thesis is part of the research project Congested railways that is included in a research program for capacity analysis and simulation at KTH Royal Institute of Technology, 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.

1.1 Objectives

The main objective of the thesis is to determine how different factors affect railway infrastructure capacity. It is possible to distinguish between theoretical capacity and practical capacity. Theoretical capacity can be achieved under ideal circumstances. In reality however, practical capacity is often significantly lower due to limitations and imperfections not considered when theoretical capacity is calculated. Theoretical capacity can be useful for long term strategic capacity planning, while practical capacity is of more interest at tactical and operational stages. Train delays play an important role in capacity analysis and are one of the main reasons why theoretical capacity cannot be achieved in practise. Consequently, a large part of the research presented in this thesis is focused on train delays.

To be able to define railway capacity, it is necessary to understand the underlying principles of railway operation and how they are affected by increased traffic load. One objective of the thesis is to determine how traffic load affects the quality of operation, e.g. scheduled running times and sensitivity to delays, and what parameters are of importance. The relationship between capacity utilization and quality of operation is important in order to determine when the system is saturated, i.e. when the consequence of adding more trains outweighs the benefits.

The timetable is of great importance in capacity analysis of structured railway operation. In Sweden and many other countries, railway infrastructure is not designed for a specific timetable and the timetable changes from year to year. The uncertainty about future timetables makes it challenging to analyse effects of infrastructure investments. This is especially true for major investments that often have a timespan of several years from the planning and designing stage until operation can start. By developing methods that can take this uncertainty into consideration, one aim of the thesis is to improve the reliability of capacity analyses of future scenarios. It is also of importance that developed methods are applicable in analyses of realistic and life sized scenarios.

1.2 Delimitations

The capacity of a railway can be defined in many ways. 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.

Effects of capacity utilisation are in this work limited to primarily SWT and delays rather than economic evaluations. However, many of the results presented in this thesis can serve as input to economic evaluations.

The analysis performed in this thesis addresses railway traffic during normal operation. With normal operation in this case means operation without too large delays. Sources of large delays, disruptions, can e.g. be complete stops caused by 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 and the timetable is no longer relevant. Consequently, it requires other methods of analysis than is employed in this thesis (Cacchiani et al. 2014).

Many of the studies presented in the thesis are based on Swedish conditions. This includes for example infrastructure design, timetable construction guidelines, train vehicle models and primary delay levels. However, the developed methods and many of the major conclusions are general.

Primary delays are modelled as independent in the simulations performed in this thesis. This is a simplification of reality. Examples of common events that generate delays that are not independent are trains operating at reduced traction power, temporary speed restrictions or headway dependent dwell times. These kinds of delays are not separated from delays of more random nature when delay distributions used in the simulation studies are compiled from data from real operation.

1.3 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. A *block section* is a section of track that can only be occupied by one train at a time. Shorter block sections give shorter minimum headway times, figure 1.2, 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. Minimum headway is the shortest time interval between two successive trains that is possible to have without the second being interfered by the first.

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.

Train operation can be either structured or unstructured. If the *structured operation*, trains are operated according to a planned timetable. In *unstructured operation* no timetable exists and train can depart whenever ready without consideration of the pre-planned timetable. Most passenger services are operated according to a timetable while it is more common with unstructured operation of freight trains, especially in the United States.

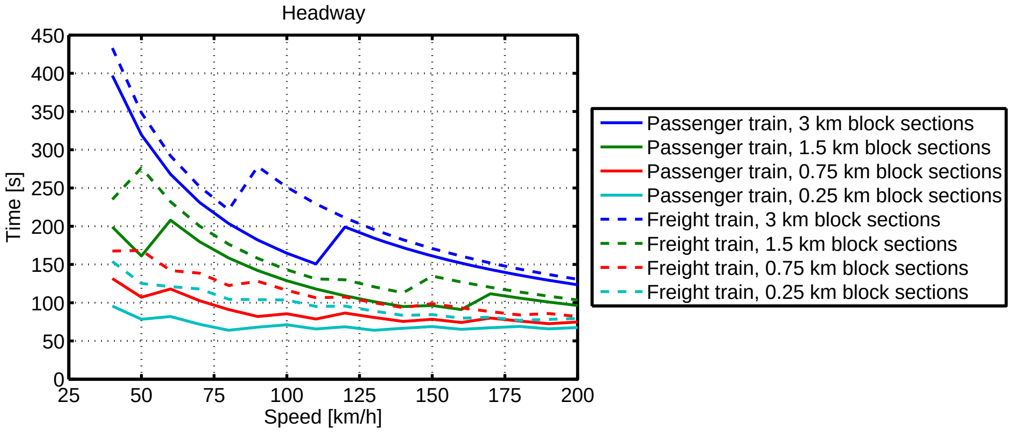


Figure 1.2: Example of minimum technical headway on a double-track section equipped with Swedish ATC2 with infill (Lindfeldt 2008).

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 of structured operation is analysed. The performance of the timetable decides the number of trains that can be scheduled and their scheduled running time, but it also affects 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. A timetable can be characterized by *static* and *dynamic* properties. Static properties describe the timetable as planned while dynamic properties describe how a timetable performs when used in operation. Examples of static properties are timetable heterogeneity, scheduled waiting time, buffer times etc. Examples of dynamic properties are train delays and lost connections between trains.

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* or *scheduled waiting time (SWT).*

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 1.2. Consequently the performance of the rolling stock in terms of acceleration and breaking performance is an important factor, both because it means higher mean speeds and shorter breaking distances, but also because it decreases the sensitivity to delays thanks to that less time is lost due to unplanned stops and speed restrictions.

Train 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 1.3 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.

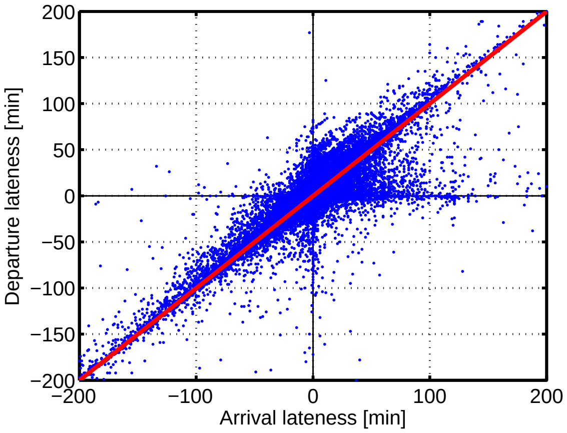


Figure 1.3: Left, observations of arrival and departure lateness at scheduled stops for some Swedish stations.
Period of measurement: September-October 2008 (Lindfeldt, 2009).

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 can be 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 the same resource at the same time and thus one train has to wait. Such resources can e.g. 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 have when it enters and leaves the analysed system.

2 Related research

There are several different methods of analysing railway operation. Different methods can be divided into analytical, combinatorial and simulation based, Mattsson (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 at all, and that the effects of dispatching are 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. In general, methods based on less detailed models may be better for drawing general conclusions. On the other hand, more detailed models are required to perform more thorough studies, but they do also require more data as input and risk generating results that are only valid for a specific setup.

2.1 Timetabling, robustness and delay estimation

Harrod (2012) gives an overview of optimisation based model structures used for railway scheduling. He classifies models according to if the track structure is modelled explicitly and if the timetable is periodic or not. Four different fundamental model structures are discussed: Mixed integer sequencing linear problems (MISLP), Binary integer occupancy programs (BIOP), Hypergraph formulation and Periodic event scheduling. The fundamental properties of each model are described together with their advantages and disadvantages.

De Fabris et al. (2014) develop a heuristic for timetable generation in large networks. Infrastructure model is mesoscopic, which makes it possible to calculate timetables fast at the same time as it is possible to consider e.g. restrictions incurred by signal blocking times and train route dependencies in station switch regions. The mesoscopic model can be used to calculate train running times and minimum headway times between trains, which is an advantage compared to macroscopic models where this have to be provided as input. The heuristic can calculate a timetable for the network in north-east Italy in a few minutes. The timetable has such quality that it is accepted by timetable planners.

Goverde and Hansen (2013) give several timetable performance indicators:

- *Infrastructure occupation.* Can be obtained by compression of timetables, UIC 406 (2004), and gives infrastructure occupation as a percentage of a time period the infrastructure is occupied by train movements.
- *Timetable feasibility.* A feasible timetable should not have any conflicts between trains, i.e. that several trains are scheduled to use the same infrastructure at the same time. Scheduled running times and dwell times should not be shorter than is possible for trains to perform in reality.
- *Timetable stability* is the ability of a timetable to recover from a primary delay without active dispatching. Can be measured as the combination of the size of the primary delay and the corresponding settling time, i.e. how long time it takes before all trains return to their scheduled train paths. Is dependent on timetable allowances and buffer times.
- *Timetable robustness.* A robust timetable can handle the variance in process times that often occur in real operation, .e.g. due to different driver behaviour and passenger volumes and weather. Both primary delays and secondary delays are minimized in a robust timetable. Is dependent on timetable allowances and buffer times.

- *Timetable resilience*. Is the possibility to avoid or minimise secondary delays using train dispatching, i.e. it is allowed to change the order of the trains in order to reduce secondary delays. Timetable resilience can be measured in train punctuality and mean delay.

The timetable performance indicators are used to define four timetable design levels: 1: Stable, 2: Feasible, 3: Robust and 4: Resilient. A higher timetable design level requires the lower levels to be fulfilled, is more challenging to achieve, requires more data and more advance modelling.

Rudolph (2003) looks into the effects 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 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.

Huisman and Boucherie (2001) 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. The work is continued by Huisman et al. (2002), where train waiting times are estimated in a railway network using a queueing model. Estimated waiting time is the sum of scheduled waiting time and delay.

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.

Salido et al. (2012) continue the work presented in Abril et al. (2008) and develop both fast analytical methods and a simple rescheduling algorithm to estimate the impact of a short disruption on secondary and total delay. Several factors are varied including traffic density, heterogeneity and allocation of running time allowance. However, primary delays are modelled in a simplified manner as only one primary delay is inserted at a time. The methods are good to establish how fast a timetable can recover from a disruption, which is indeed one definition of robustness, but is not designed to analyse real life operation with multiple disturbances.

Deadlocks occur in situations when several trains are blocking each other so that train movements cannot continue. Problems with deadlocks are present in synchronous simulation of train operation, while it is not a problem in asynchronous simulation and situations of rescheduling (Pachl 2007). He further illustrates the problem of deadlocks and he gives two different approaches for avoiding deadlocks that can be implemented in simulation software: Movement Consequence Analysis and Dynamic Route Avoidance. Deadlocks are dependent of infrastructure layout and occur on track sections with bidirectional traffic, e.g. single track railway lines. It is emphasized that it is a necessity that proposed methods for deadlock avoidance must not be complicated and computational intense in order to avoid long simulation times. It is a process that has to be continuously ongoing in the simulation since trains might be disturbed by primary delays at any moment. Hence, methods applied for rescheduling is not applicable.

Andersson et al. (2015) use a MILP model to increase the robustness of an existing timetable. The model is an extended version of the one presented in Törnquist and Persson (2007). The robustness of the timetable is calculated by identifying critical points, RCP (Andersson et al. 2013), where the risk for train interaction in case of disturbances is high. Critical points occur in locations where train systems enter the line and overtakings are scheduled. They are calculated pairwise between two trains and are composed of three parts; running time allowance before the critical point, running time allowance after the critical point and headway between the two trains in the critical point. After all critical points have been identified, the model is used to calculate a new timetable by redistributing existing running time allowance and shift trains in the timetable, without changing train order, to increase the robustness in the critical points identified. That the robustness has indeed increased is verified by reusing the optimisation model to simulate a real-time rescheduling where trains are delayed by entry delays. Compared to the original timetable, train delays are smaller in the modified timetables where RCP values have been increased.

In a similar study, Khoshniyat and Peterson (2015) use the same MILP model to study the effect of travel time dependent minimum headways on train delays. The underlying assumption is that trains accumulate delays as they run from origin to destination, which was shown in an empirical study performed by Peterson (2012). An existing timetable is adjusted so that headways increase with distance travelled, where they are more efficiently used to avoid secondary delays. Three different delay scenarios are used, delay of a single train at a random location, speed reduction of a single train (the entire trip) and speed reduction for all trains passing a specific location. It is concluded that delays do indeed decrease when headways are redistributed and increased towards the end of the train runs.

Medossi et al. (2011) introduces stochastic blocking times instead of deterministic times that are normally used in blocking time models. Stochastic blocking times make it easier to determine the probability that two trains will interfere with each other and cause secondary delays. GPS data is used to analyse train run characteristics in detail, including acceleration, cruising, coasting and braking. For each motion phase, separate performance parameters are estimated, including e.g. reduced rate of acceleration, cruising speed, coasting time interval and braking rates in different situations. The variability in the estimated parameters are used to create the stochastic blocking times, but does also reveal interesting details about driver behaviour. The proposed method can also be used to provide input to and calibrate micro-simulation tools. The stochastic blocking times are estimated for each single train run individually but do not include effects due to secondary delays.

Cerreto (2015) proposes a micro simulation based method for estimation of timetable robustness. Primary delays are applied to a single train in the timetable and the consequences in terms of e.g. settling time and the number of delayed trains is measured. Which train that receives the primary delay is systematically varied and the simulation process repeated. A technique is also presented that reduce the large number of simulations needed.

The differences between and limitations of micro, meso and macroscopic infrastructure models are discussed in Gille et al. (2008) and one of the conclusions is that station areas should preferably be modelled in higher detail than at macroscopic level. Cui and Martin (2011) argues for the benefits of simulation models that can combine all three levels of detail in a flexible way. The balance between accuracy and performance, i.e. computational workload, should be considered when the model is constructed. A microscopic infrastructure can be evaluated according to certain criteria in order to determine the importance of different components. Important components can remain at microscopic level while less important components can be aggregated to be mesoscopic or macroscopic. The lover level of detail of appropriate parts of the infrastructure model makes computation faster without losing accuracy.

Büker and Seybold (2012) describe an analytical model for analysis of delay propagation in railway timetables. Delays are modelled by cumulative distribution functions and train interaction calculated by means of an activity graph. A challenge that is addressed in the paper is to find a suitable class of distribution functions that can be used to model delays realistically and can be used in the analytical calculations. The analytical approach makes the model faster than running simulations. Dispatching decisions is modelled in a simplified manner and higher-order second secondary delays are neglected. However, the effects of the latter are considered to be small in real world scenarios. The methodology is implemented in a software tool. Another analytical method for timetable stability analysis of cyclic timetables using max-plus algebra is presented in Goverde (2007).

Kroon et al. (2008) use stochastic optimisation to improve robustness of an existing timetable by reallocating buffer times and running time allowances with the objective of minimising train delays. It is applied to cyclic timetables, but can be used on non-cyclic as well. However, cyclicity reduces computing time. The method is a two stage recourse model that includes a timetabling part and a simulation part that evaluates the robustness of the timetable. The simulation part of the model does not include dispatching functionality, i.e. it assumes that trains always run in the same order as in the scheduled timetable. Timetables optimised by the model have less delay compared to the original timetable, both in simulation and real world experiments. An example of another model that can create optimal timetables with respect to both train running times and timetable robustness is presented in Fischetti (2009).

2.2 Capacity

Abril et al. (2008) give a good introduction to the concept of capacity and how it is affected by different properties including infrastructure, traffic and operating parameters. They also give different methods how it can be calculated, including analytical methods, optimization methods and simulation methods. They state that while analytical and optimizing methods can be used to estimate theoretical capacity, simulation is an attractive way to obtain practical capacity measures where the trade-off between capacity and reliability, i.e. delays, can be included.

Pouryousef et al. (2015) present a review of different methodologies for capacity evaluation used in Europe and the U.S. based on more than 50 different capacity studies. Differences between Europe and the U.S. railways are highlighted and one conclusion is that these differences affect how capacity is measured and what tools are used. One of the most important differences is that freight traffic dominates in the U.S. while passenger traffic dominates in Europe. Another important difference is that that trains are normally operated according to a timetable in Europe, structure operation, while this is not the case in the U.S. where normally no detailed timetable exists and conflicts are solved by improvising, improvised operation. For this reason, capacity analysis in Europe tends to be more timetable oriented, e.g. application of the UIC 406 model or timetable-based simulation.

Sameni (2011) concludes that there are three main groups of capacity measures: throughput (e.g. number of trains), Level of Service (e.g. delay) and asset utilisation (e.g. velocity and infrastructure occupation time).

White (2005) discusses the importance of how simulation results are evaluated. Several different measures of level of utility that is commonly used to evaluate infrastructure improvements are described and their advantages and disadvantages are examined. Several scenarios are compared in a case study and it is shown that the conclusions are different depending on what measures are used to evaluate the results. While non-North American studies consider traffic management decisions, it is not as common in North American studies, which may lead to false conclusions. It is also suggested that statistical analysis of simulation results are, if possible, complemented with a root cause analysis that can be used to identify locations where infrastructure investments are needed and to validate results.

In UIC (1996) capacity is defined by three key parameters, Infrastructure, Operating plan and Quality. SWD and delays are both included in “Quality” and are important to consider when capacity is analysed. Capacity can only be evaluated with respect to an operating plan, where timetables are an important part. 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.

Lindner (2011) discusses the applicability of the UIC code 406 (2004) method for evaluating 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 complemented 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 are addressed in the 2nd version of the UIC 406 capacity leaflet (2013).

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 (Scwanhäußer, 1974) 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.

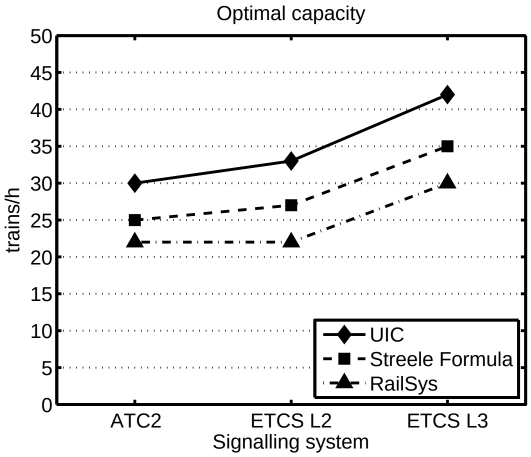


Figure 2.1: Estimated capacity (trains/h) of the bottleneck in Stockholm using different methods (Magnarini, 2010).

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 double-track railway line. The investigated parameters are: distance between adjacent overtaking stations, train top speed, train frequency, entry delays and running time extensions, for both high-speed services and freight services independently. In order to reduce the number of necessary parameters, the delays are modelled by negative exponential distributions. In addition, Lindfeldt points out the difficulty of defining the timetable by a few independent factors. Thanks to the experimental design using Latin hypercubes, only 66 design points are needed to form the metamodels. The simulations are performed in the simulation tool RailSys using the mean and standard deviation of the delays as response variables. The results show that the speed and frequency factors as well as the running time extension have great impact on delays. The entry delays and inter-station distance are found to have less impact.

In Lindfeldt (2009) a combinatorial model, TVEM, is developed. By varying starting times of the different train patterns, a large set of timetables are created using a sequential scheduler. The first steps are to schedule cyclic train services, normally passenger trains. In the final stage, the remaining capacity is used to schedule as many train paths as possible, normally freight trains. The quality of the train paths is controlled by limiting the scheduled waiting time. A similar approach is presented in Meng (2013) where a rolling horizon optimization algorithm is used to measure capacity by iteratively schedule more freight trains in a timetable with passenger trains.

Sipilä (2014) uses RailSys to generate and simulate timetables on a single track railway line. There is no native support in RailSys to quickly generate several timetables. However, the paper presents a method where RailSys simulations are used to generate timetables. This is achieved by systematically designing entry delays in such a way that they give trains their desired departure times. No other primary delays are applied in the simulation and the results are then used as timetables in a second step where they are simulated with stochastic primary delays. The framework is further developed in Sipilä (2015) to include automatic infrastructure generation.

Landex (2008) discusses a simulation model for strategic capacity analysis, SCAN, developed in Kaas (1998). The model uses discrete synchronous simulation to generate conflict free random regular interval timetables for a railway network with mesoscopic infrastructure representation. A plan of operation is used to generate several timetables that are evaluated. The timetables are ranked with respect to SWT and the 25th percentile is used as a measure for quality of operation. The timetable with the lowest SWT is not used because it may not be representative for the timetable preferred in practice. This is due to factors not included in the evaluation, e.g. passenger transfers and bundling of trains. SCAN is not used for operational simulation of timetables with applied disturbances, hence it is not used to evaluate delays.

Eliasson et al. (2014) emphasize the importance of the timetable in railway investment appraisals. Timetables are required to estimate the social benefit of railway investments and explicit principles are needed for the assumptions involved in the analysis or results will be arbitrary and scenarios incomparable. This is especially important for capacity improvements which can be used to improve frequency of service, travel time or reliability. Preferably, social benefits of using different timetables in the do-nothing and investment scenarios should be systematically investigated. It is emphasised that the performance of the timetable in the do-nothing scenario and investment scenario are equally important.

Burdett and Kozan (2005) calculates absolute capacity of railway networks using optimisation to find bottlenecks in sections/lines, hence it is called a bottleneck approach. Each line has a bottleneck that limits capacity, and by finding the capacity of that bottleneck, the capacity of the whole line is determined. The presented work improves the bottleneck approach by including several important factors of railway operations, such as e.g. train mix, dwell times and more realistic headway modelling. The method can be used to determine where bottlenecks are located in the network, which can be useful for planning of infrastructure investments. It can be used to analyse networks, rather than single lines, and can therefore be used to analyse interaction effects between several interconnected lines. One of the conclusions is that railway lines cannot be used efficiently when they are part of a network. Train delays and timetable robustness are not included in the analysis.

Wittrup et al. (2015) presents a new model for capacity consumption calculation. It is similar to the UIC 406 method as it calculates capacity consumption by analysing how much timetables can be compressed in time. However, contrary to the UIC method, the proposed model does not require a timetable as input and it can also evaluate capacity under stochastic conditions, i.e. when trains are affected by primary delays. An operating plan is used to generate all possible permutations of train sequences, or a sample if the total number is too large, and an asynchronous scheduled is used to schedule trains based on a first-in first-out principle. In a second step, scheduled train sequences can be simulated with primary delays of type entry delay and dwell time extensions. Different train sequences give different results, hence results are presented as distributions of capacity consumption. The method can be used to analyse capacity in networks or routes. One conclusion is that capacity diminishes when a network is analysed, compared if just a single route or line is included in the model. A second conclusion from the case study is that measured capacity increase from infrastructure investments is greater if delays are considered in the evaluation. Future work includes possibility for trains to overtake other trains and other couplings between trains.

Lai et al. (2013) develop a railway capacity model inspired by an approach used in the highway capacity manual to analysis capacity of road traffic. A new concept of base train equivalents (BTE) is introduced that can conveniently be used to analyse capacity of railway lines under different conditions. All train types are converted to BTE using a headway based approach where e.g. train performance, dwell times, properties of the signalling system as well as the combination of preceding and following train type is considered. Hence different train types will have different BTE values. The model is validated against an existing capacity model.

Goverde et al. (2013) compare two different signalling systems, Dutch NS’54/ATB and ETCS Level 2, and their influence in capacity. The proposed method expands the standard UIC compressions method for capacity consumption calculation with a dynamic part that calculates capacity consumption and delays in disturbed conditions. Trains are disturbed using stochastic entry delays and train paths simulated, by means of Monte Carlo simulation, using the train dispatching system ROMA that models the infrastructure at a microscopic level, (D’Ariano and Pranzo, 2009). A case study is performed on a 40 km long double-track line using a timetable from real operation (2 hour morning peak). Measures of performance include scheduled infrastructure occupation, dynamic infrastructure occupation, average delay, secondary delays and punctuality. Several different strategies for train dispatching in disturbed conditions are tested and results show that strategies that yield low dynamic infrastructure occupation gives higher delays and vice versa. Hence, one of the conclusions is that it is necessary to consider both dynamic infrastructure occupation and delays in the analysis. Still to be implemented are primary delays of type dwell time extensions and running time extensions.

Yuan and Hansen (2007) present an analytical stochastic model for capacity evaluation of stations. The model estimates secondary delays due to route conflicts and train connections while taking constraints imposed by e.g. the signalling system into account. It is possible to determine the maximum frequency of trains given a maximum accepted amount of secondary delays. One conclusion is that when scheduled buffer time between trains decreases, secondary delays increase exponentially.

Murali et al. (2010) develop a simulation-based technique to generate delays used in regression models to predict delays in double- and single-track lines. Several parameters are used to describe the topology of the network as well as the operating conditions when the train of interest enters the subnetwork. They find an exponential relationship between delay, train mix and parameters describing the operating conditions and network topology.

Yung-Cheng and Yung-An (2012) create parametric models to estimate capacity of single and double-track operation. The microscopic simulation software Rail Traffic Controller (RTC) is used to perform a full factorial design. Traffic consists of freight trains that are operated without a timetable and stochastic entry delays are applied in the simulations. Output from the simulation, train delays, is used to estimate parameters in both regression models and in a neural network (NN) model. Factors in the model for single track operation are siding spacing, signal spacing, track speed, volume (trains/day) and heterogeneity. For double-track, the factors are crossover spacing, signal spacing track speed, volume, and heterogeneity. A measure of heterogeneity is defined that is applicable if the traffic mix consists of two types of trains. The conclusion is that the regression model performs better in estimating single track operation while NN is better on double-track operation.

2.2.1 Heterogeneity

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. Overtakes are also possible. The measures are further developed in Vromans (2005) by compensating for the minimal headway that is technically possible between two trains at the each location, thereby estimating the headway buffers rather than the absolute size of the headways. This has an advantage if the minimum technical headway varies along the line or between different train types. The SSHR and SAHR are further developed by Landex (2008) into new measures for heterogeneity that is independent of traffic density and number of trains used in the calculation.

Sogin et al. (2011) analyse the effect of heterogeneous traffic on a single track freight network. The analysis is performed with RTC, and the measure of performance is delay of the freight trains in min per 100 train miles. Delay includes the time needed for meets and passes, i.e. they are not planned in advance. Traffic density is varied and heterogeneity is controlled by systematically adding passenger trains of different speeds. For completely homogenous 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 Sogin et al. (2013) the method is further developed and used to analyse how incremental transition from single to double-track affects train delay. Simulation results are used to construct a response surface model and it is concluded that train delays decrease linearly with additional sections of double-track. The reduction in delay is greater for higher traffic volumes.

Gibson et al (2002) develops a regression model using delay data from the rail network in the UK. He uses a method similar to the timetable compression defined in the UIC 406 (2004) leaflet to define capacity utilisation. He tests a number of functional forms (exponential, adjusted exponential, power and linear), and finds that secondary delays increase exponentially with capacity consumption on a line section. He also discusses how the relative speed of a train affects its marginal cost, congestion cost. Using simulation, he concludes that adding a train that is 20 % faster than the fastest train in the timetable, have a congestion cost that is 20 % higher than predicted for a train of average speed. Similarly, adding a train that is 20 % slower, costs 50 % more than the average train.

Harrod (2009) uses a hypergraph model to analyse the effect of introducing a fast passenger service on a single-track railway line with homogenous freight traffic. The effect of introducing the passenger service is measured in loss of utility for the freight trains. Utility include the number of freight trains it is possible to schedule and their scheduled waiting time. If a freight train path has a negative utility, e.g. due to too much scheduled delay, it is cancelled and the freed capacity can be used by the other trains to get better paths. Three basic layouts of the infrastructure are analysed, single-track with evenly spaced two-track stations, single-track with evenly spaced three-track stations and a single-track with two-track stations where the middle third is replaced by double-track. Stations are as long as the line sections between the stations. Passenger trains have absolute priority over freight trains and two different speed levels are analysed, 33% and 100% increased speed compared to the freight trains. Results show that if passenger trains run faster, the negative effect imposed on the freight trains will be smaller. They also suggest that it is more efficient to upgrade stations to three tracks rather than constructing a long double-track section.

2.3 Comments

This thesis is about railway capacity. It is focused on analysing effects related to capacity utilisation in structured railway operation and rest on several of the areas of research that is covered in the literature review.

The timetable is essential when structured operation is analysed. Hence, a natural first step is to perform some kind of timetable analysis. This implies that you either have to analyse an existing timetable or create a timetable that can then be analysed. General capacity relationships can be established by changing conditions and making new timetables. Methods for timetable generation covered in literature include both simulation and analytical models. Constructing timetables are time-consuming and some work applies timetable-free techniques for capacity estimation. Paper I analyse an existing timetable and paper II-VI applies simulation for timetable construction.

Capacity relationships in this thesis are primarily studied by analysing timetables. A timetable has many properties. So called static properties are straight forward to evaluate analytically, such as e.g. scheduled waiting times, headways between trains and heterogeneity. Dynamic properties involving how the timetable will behave when it is used in operation is harder to derive. Both static and dynamic properties are of importance in capacity analysis.

Methods in literature used for estimation of dynamic timetable properties include several different analytical methods and measures, calculation of capacity consumption by compression of timetables, optimisation algorithms for train rescheduling and timetable simulation. In paper II-VI the relationship between capacity utilisation and train delays is studied. Simulation is employed since it is a method that makes it possible to calculate train delays under realistic conditions. Other methods do either not calculate train delays explicitly, or model train dispatching decisions or primary delays in a simplified manner. Modelling of primary delays and dispatching is central when train delays are analysed, Siefer (2008).

Measuring train delays, rather than some measure of robustness, have several advantages because it is what is actually experienced by the customers. Consequently many policy rules regarding practical capacity are connected to delays, e.g. that operation should meet minimum requirements on punctuality or mean delay. Furthermore, delays have a monetary value that can be used to compare it to other timetable properties such as e.g. SWT and frequency of service in socio economic evaluations. However, a disadvantage with timetable simulation is that it is time-consuming and few simulation studies in literature analyse capacity without being restricted to one or very few timetable scenarios. Methods are developed in paper II and V that speed up the process of timetable simulation, which makes it possible to simulate a larger number of timetables and scenarios.

3 Performance evaluation using empirical data

3.1 Method

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 it will reflect reality and not be affected by model limitations. Of course, it 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 paper I aims to describe the whole Swedish railway network. Data from four different databases supplied by the Swedish Transport Administration is compiled into descriptive parameters covering different aspects of the 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.

3.1.1 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. 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.

- TFÖR/LUPP (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 TFÖR contains errors, e.g. missing train movements etc. 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.

3.1.2 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 3.1.

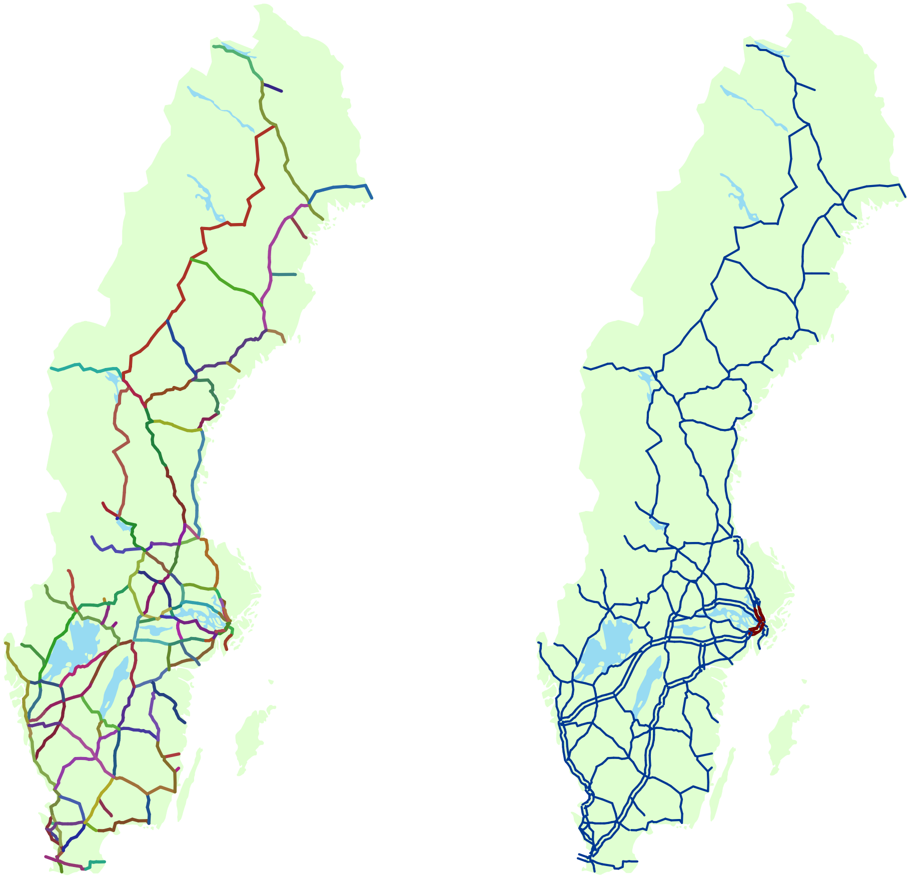


Figure 3.1: Left, line sections of the Swedish railway network used in the analysis. Right: Single track, double-track and quadruple track (red). From 2008.

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 3.1 below.

Table 3.1: Performance indicators.

Infrastructure	***Timetable***		***Traffic***	***Delays***
Single track:	**No. Trains per day**		**Freight trains**	**Passenger/Freight**
Distance between crossing stations (km)	**Total/Passenger/Freight**			
min	No. Trains per day		**Weight (metric tons)**	**Proportion of trains with increased delay**
max			min	
mean	**No. Trains per hour**		max	**Median of increased delay normalized by**
standard deviation	**Total/Passenger/Freight**		mean	**route distance [min/100 km]**
	Peak hour	Time for peak hour	standard deviation	
Proportion of stations with more than 2 tracks	Morning	06-09		**Standard deviation of increased delay**
	Afternoon	15-18	**Length (m)**	**normalized by route distance [min/100 km]**
Proportion of trains with simultaneous	Afternoon	16-17	min	
entry capability	Day	9-15 och 18-20	max	
	Night	20-06	mean	
Double track:			standard deviation	
Distance between passing stations (km)	**Speed (km/h)**			
min	**Passenger/Freight**		**No. Axles**	
max	max		min	
mean	min		max	
standard deviation	mean		mean	
	median		standard deviation	
All lines and stations (m)				
Track length	**Speed difference**		**Axle load**	
min	standard deviation		min	
mean	standard deviation/mean		max	
max	95 percentile/10 percentile		mean	
			standard deviation	
			Gross tons/day (metric tons)	
			Passenger trains	
			Proportion with ≤ 12 axles	
			Proportion with > 12 axles	
Source:	**Source:**		**Source:**	**Source:**
BIS	T08.3		BANSTAT	TFÖR
Measurement period	**Measurement period**		**Measurement period**	**Measurement period**
2008-12-19	2008-10-09		2008-10	2008-09 och 2008-10

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 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 3.2 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.

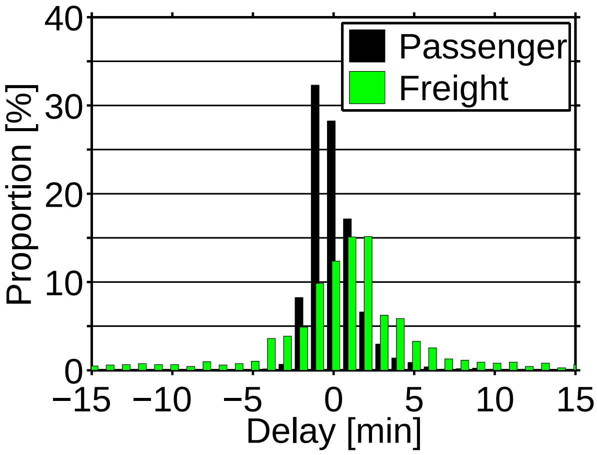


Figure 3.2: Example of registered delay development on a section of the Southern Main Line between Mjölby and Tranås (Sipilä, 2012). Distance of approximately 37 km. Period of measurement: 107 days.

3.2 Results

3.2.1 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 3.3 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.

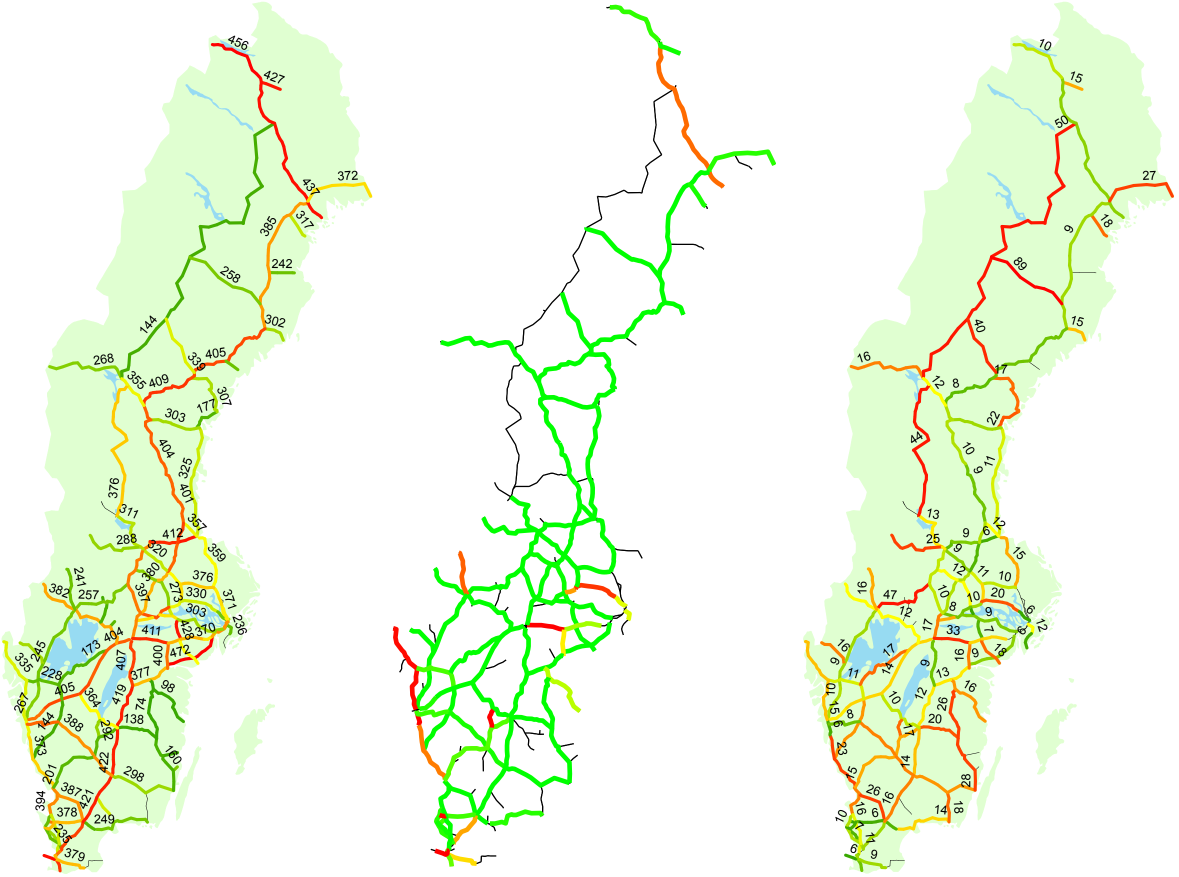


Figure 3.3: Left: Mean train length [m]. Middle: Proportion of freight trains longer than the mean track length, 0-61% (green - red). Right: Mean inter-station distance [km].

Lengths of side tracks may be a constraint for operation of long freight trains. The figure to the left 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 when stations short tracks cannot be used to meet or pass long freight trains. 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 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 Gällivare and Luleå, part of the red coloured route in the far north, where long ore-trains can only meet at five crossing stations for a distance of 204 km.

3.2.2 Timetable

The map to the left in figure 3.4 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 Hallsberg connecting Gothenburg and Malmö with the northern parts of the country. Passenger trains dominate around the larger cities Stockholm, Gothenburg 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.

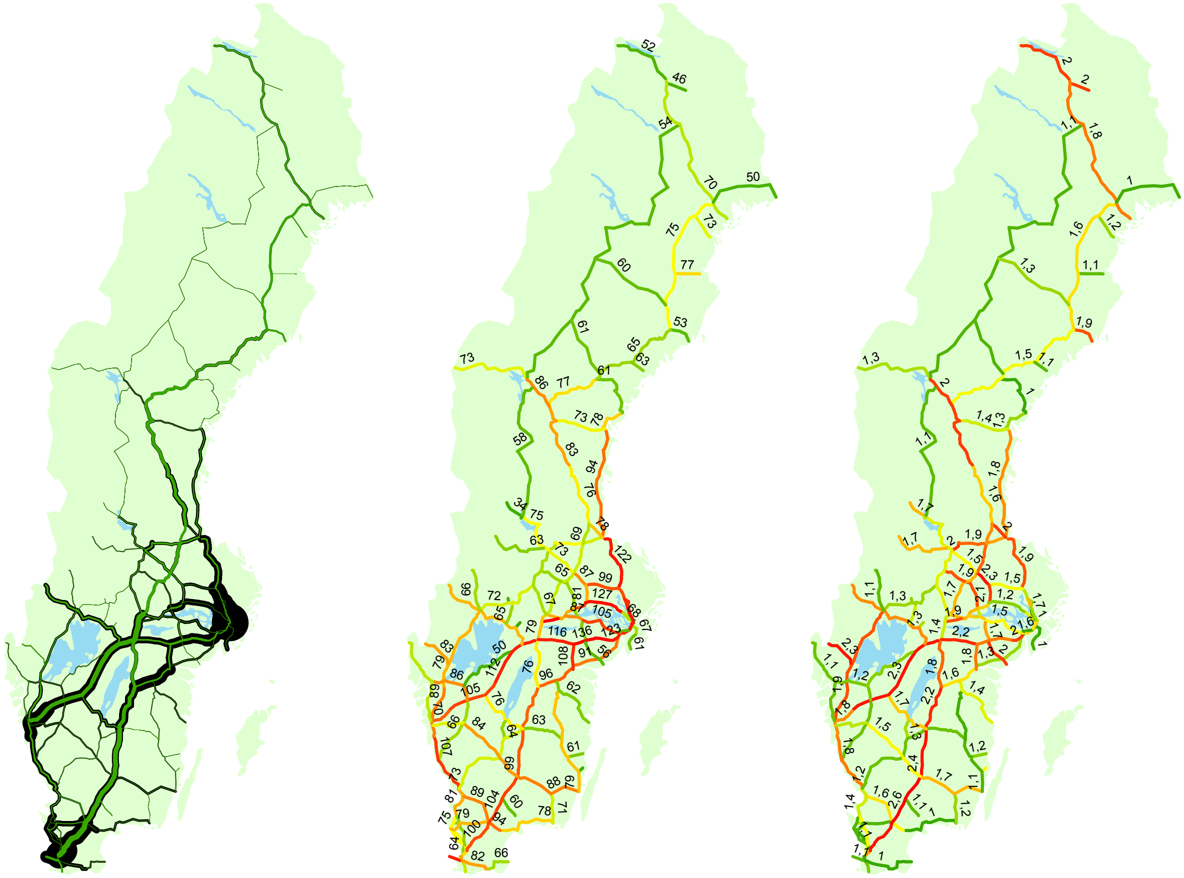


Figure 3.4: Left: Trains/day [-], black lines show the total number of trains and green lines the number of freight trains (superimposed). Middle: mean speed all trains [km/h]. Right: Speed ratio 0.95/0.10 percentile [-].

The middle figure shows the mean speed based on all trains and the right figure a heterogeneity measure. In general double-track lines have higher speeds than single track lines. This is due to both higher standard in general, but also because it is not necessary stop for crossings. The heterogeneity measure shown is the 95 percentile divided by the 10 percentile, i.e. the quota of the mean speed for a fast train and a slow train, where a high value indicates a heterogeneous timetable and 1 a completely homogenous. A similar speed ratio is presented by Krueger (1999) as the quota of the maximum speed and minimum speed. However, by using the 95 and 10 percentile, extremes that might not be representative are avoided, e.g. a freight train that have a long scheduled stop somewhere on the section. Double-track lines with a mix of passenger trains and freight trains have the most heterogeneous traffic. This is the case for the southern parts of WML and SML, where the heterogeneity value is higher than 2. The dense traffic on these lines increases the speed differences even more when freight trains have to stand aside for frequent overtakings. The heterogeneity together with the number of trains operated per day gives a hint of the capacity utilisation on the double-track lines, keeping in mind that especially freight trains and passenger trains might run during different periods of the day.

3.2.3 Delays

The left part of figure 3.5 shows the proportion of passenger trains that have increased their delay. 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 3.2. 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.

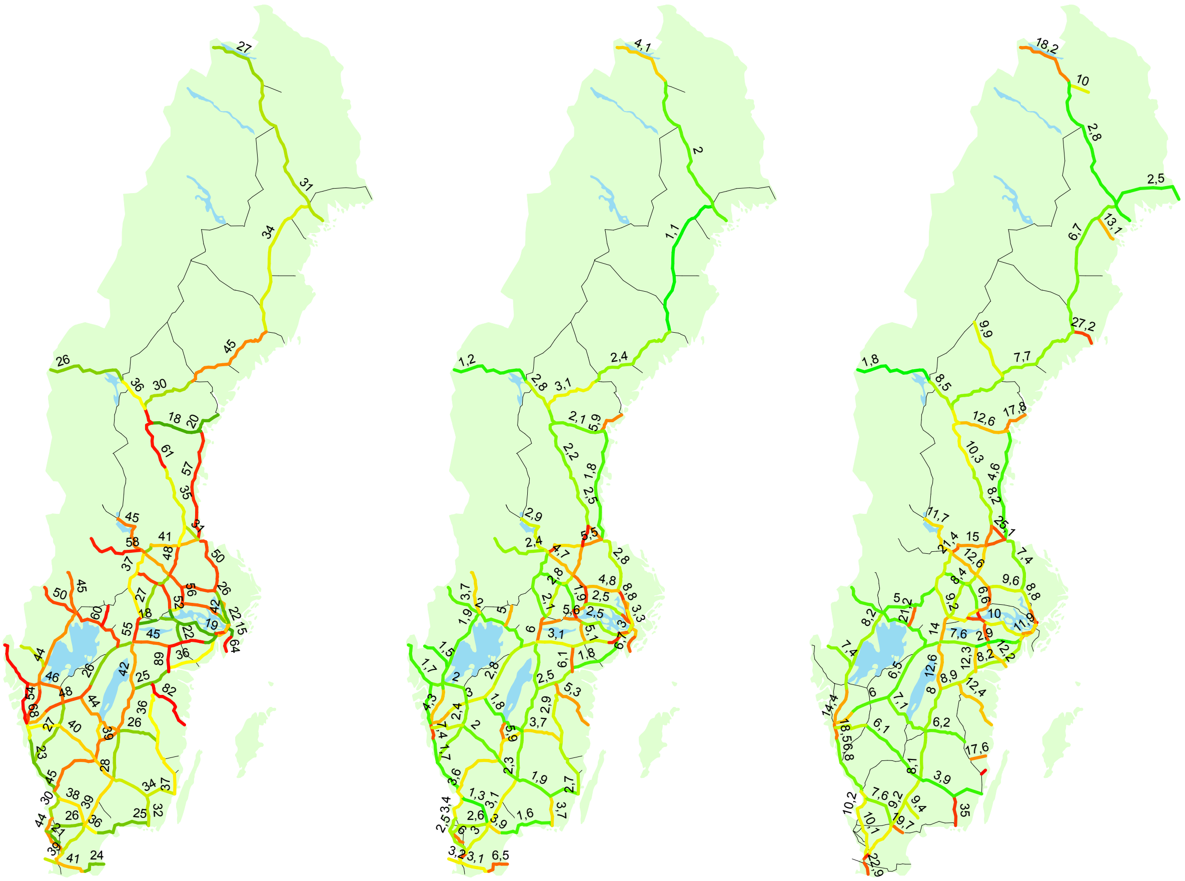


Figure 3.5: Left, proportion passenger trains with increased delay [%].
Middle, median of the increased delay for passenger trains [min/100km].
Right, median of the increased delay for freight trains [min/100km].

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 number of 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 (WML) close to Gothenburg and Stockholm. Approaching the cities, the proportion of delayed trains is 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.

3.2.4 Delay correlated parameters

The performance indicators are defined with the intention to describe properties associated with capacity. Infrastructure indicators indicate available capacity and timetable and traffic indicators use of capacity. Based on the assumption that delays are correlated to capacity utilisation, stepwise multilinear regression is used to find out how the calculated parameters correlated to the delay parameters. The algorithm uses p-values for the F-statistic to decide which parameters to include in the model. The analysis is separated for single/double-tracks, passenger/freight trains and for the proportion of delayed trains and their median delay. To refine the test, routes with extreme conditions, such as routes with extremely low traffic load, temporary speed restrictions, partial double-tracks (single-track routes), multiple-tracks or very short routes, were discarded. 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 3.2.

Table 3.2: Parameter analysis (single tracks), significant at the 0.05 level. The table shows the slope for each parameter included in the linear model.

Train type	Delay type	Parameter	Slope
Passenger	Proportion	Total nr of trains/day	0.44
Passenger	Proportion	Mean speed (all trains)	0.42
Passenger	Proportion	Speed mix (all trains)	0.29
Passenger	Median	Route length	-0.58
Passenger	Median	Share of passenger trains with > 12 axles	0.24
Passenger	Median	Std of the inter-station distance	-0.41
Freight	Proportion	Total nr of trains/day	0.57
Freight	Proportion	Share of stations with at least 3 tracks	-0.46
Freight	Median	Route length	-0.72
Freight	Median	Freight train mass	0.37
Freight	Median	Max inter-station distance	-0.37

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 3.6.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.

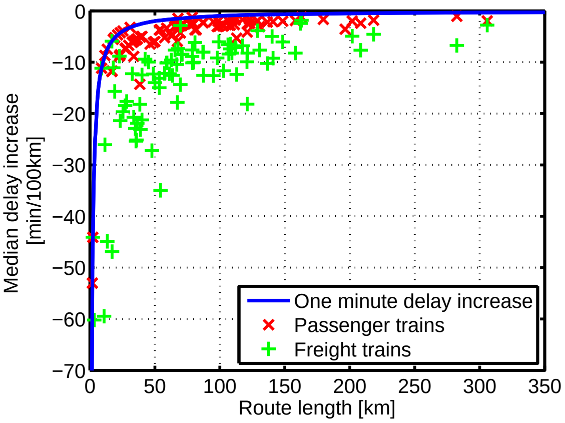


Figure 3.6: The solid line shows the effect of the route length of the minimum possible delay increase.

3.3 Conclusions

The performance indicators are good to describe the current status of the railway network, including important infrastructure characteristics and how the network is utilised. Presenting results on maps of the national network enables a macroscopic perspective which makes it possible to quickly identify weak spots in the system.

Except for the division of the network into line sections, the developed method is completely automated. This makes it easy redo the calculations with new updated data. The study from 2008 was followed up by a study based on data from 2012 (Lindfeldt 2014). It is focused on the development between 2008 and 2012. Some of the conclusions are that traffic has increased between 10-40% on many line sections while there was no corresponding increase in delays. However, the average speed had dropped 5-10% in many sections, which might be an effect of increased congestion or that more allowances are added to the train running times.

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. The data used in the analyses did not contain any information about the allowances in the timetables. Allowances are used to catch up delays and can compensate for increased secondary delays. Therefore it can be hard to correlate capacity utilization and delays without knowledge of the allowances, especially since they tend to increase with higher capacity utilization. In general, more SWT, which can be used as allowance, are needed to construct feasible timetables when capacity utilization is high. Also, adding more allowances can be an active measure taken by timetable planners to limit the effect of delays in congested areas. This motivates why both delays and SWT has to be included when the performance of a railway system is evaluated.

4 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. The commercial simulation software RailSys is used in this thesis. RailSys is an advanced tool for timetable planning and simulation of train operation (Radtke and Hauptman, 2004). 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.

4.1 Method

To handle experiments with many scenarios, an interface is required to handle input and output from RailSys. In paper II-IV, this is done by transferring data using xml files. It is used to export results from running time calculations in RailSys and information about the infrastructure model. The data is then used by a sequential simulation algorithm to create conflict-free simulation-ready timetables. The timetable and perturbation data is then imported into RailSys for simulation. The framework is retrospectively named SAMO.

The ability to generate and import timetables to RailSys opens up several possibilities.

- To perform factor analysis of parameters influencing the timetable and perturbations. If more than a few parameters are 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.

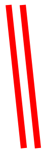
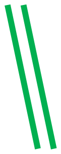
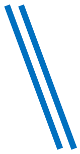
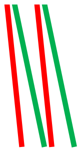
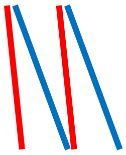
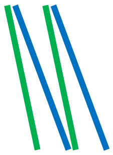
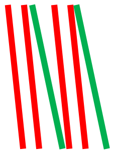
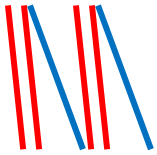
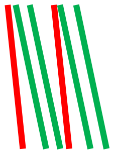
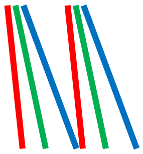
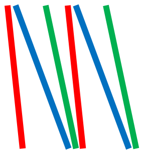
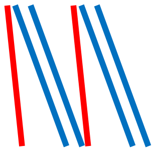
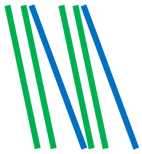
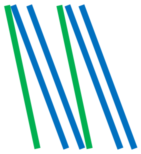
- The detailed information about the applied primary delays allows for more detailed analysis of the simulation results and can for example be used for distinguishing secondary delays from primary delays.

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. Overtaking stations with two tracks are spaced equidistantly and timetables are defined as cyclic timetables of up to three trains per cycle, table 4.2. In the scheduling algorithm, timetables of different traffic density are created by changing the headway between trains starting at the origin. Trains are then scheduled as fast as possible from origin to destination. 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 real operation (Nelldal 2008).

Table 4.1: Experimental setup.

Inter-station distance	**Timetable variant (mix of train types)**		**Traffic intensity**	**Perturbation level**
[km]	Variant number	HS: high-speed, IC: intercity, FR: freight	Headway, % of minimum	Low, high
	1	100% HS		
	2	100% IC	100	
20	3	100% FR		
	4	50% HS, 50% IC	116	
	5	50% HS, 50% FR		Low
	6	50% IC, 50% FR	138	
30	7	67% HS, 33% IC		
	8	67% HS, 33% FR	171	
	9	33% HS, 67% IC		
	10	33% HS, 33% IC, 33% FR	223	High
40	11	33% HS, 33% FR, 33% IC		
	12	33% HS, 67% FR	322	
	13	67% IC, 33% FR		
	14	33% IC, 67% FR		

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]

 1 (1) [0]	 2 (2) [0]	 3 (3) [0]	 4 (1 2) [9.5]	 5 (1 3) [27.5]
 6 (2 3) [18]	 7 (1 1 2) [6]	 8 (1 1 3) [18.5]	 9 (1 2 2) [6]	 10 (1 2 3) [18.5]
 11 (1 3 2) [18.5]	 12 (1 3 3) [18.5]	 13 (2 2 3) [12]	 14 (2 3 3) [12]	

In the experiment, explanatory variables are the inter-station distance, traffic mix (heterogeneity), number of trains per hour and level of primary delays. The dependent variables are scheduled delay, secondary delay and used allowance. Scheduled delay is a property of the timetable and easily calculated while simulation has to be used to establish secondary delays and how allowances in the timetable are used. 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. An overview of the workflow is shown in figure 4.0.

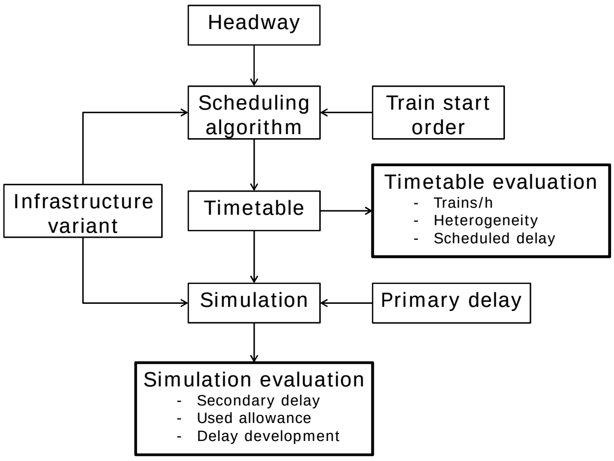


Figure 4.0: Workflow of the experiment.

4.1.1 Calculation of secondary delays and used allowance from simulation results

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 practise it becomes more complicated. Consider the example in figure 4.1. 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, illustrated in figure 4.1. 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 on the next page.

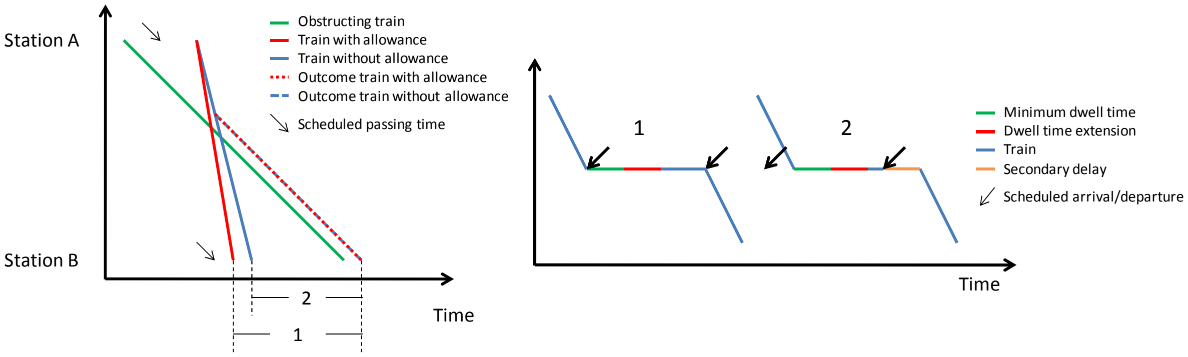


Figure 4.1: Illustration of secondary delays on line sections and at stations.

$$\begin{aligned} ual_i &= \max(\min(dd_i, al_i), \max(dd_i - da_{i+1} + re_i, 0)) \\ sdl_i &= \max(da_{i+1} - dd_i + \min(dd_i, al_i) - re_i, 0) \\ uas_i &= \min(da_i + de_i, ss_i - ms_i) \\ sds_i &= dd_i + uas_i - da_i - de_i \end{aligned}$$

ual: used allowance on line section
al: available allowance on line section
uas: used allowance on station
sdl: secondary delay on line section
sds: secondary delay on station
dd: departure delay
da: arrival delay
re: running time extension
de: dwell time extension
ss: scheduled stop time
ms: minimum stop time

4.2 Results

The bar graph in figure 4.2 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 overtakings 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 (2.3-5.3 trains/h), 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.

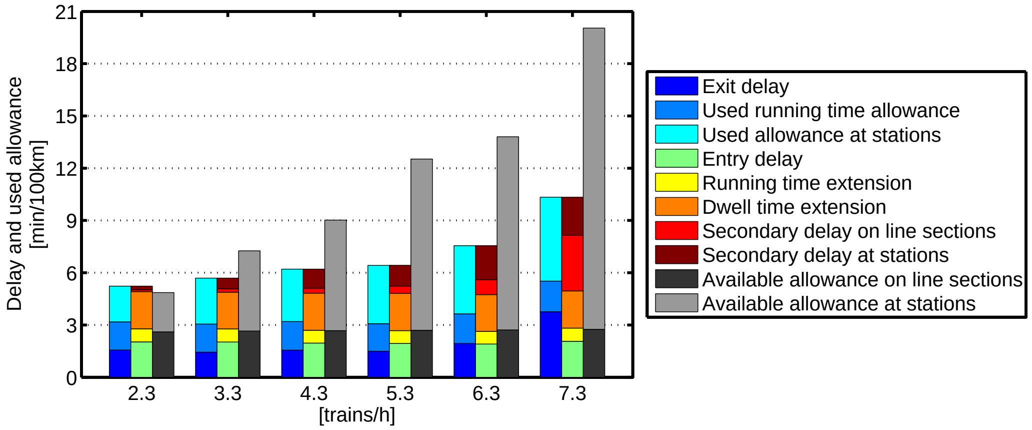


Figure 4.2: Timetable type 10 (high-speed, intercity, freight trains), 20 km interstation-distance and high perturbation level. Results are mean for all trains.

One methodological aspect apparent from figure 4.2 is that all types of primary delays are as almost 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 (green, yellow, orange, red, brown) sums up to the same values as and used allowance and exit delay. It shows that definitions of secondary delays are consistent with the definitions of used allowance.

4.2.1 Heterogeneity

One of the main objectives of paper II, is to analyse how heterogeneity affects delays and capacity. Heterogeneity 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 4.3 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 4.2). 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, most of the 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.

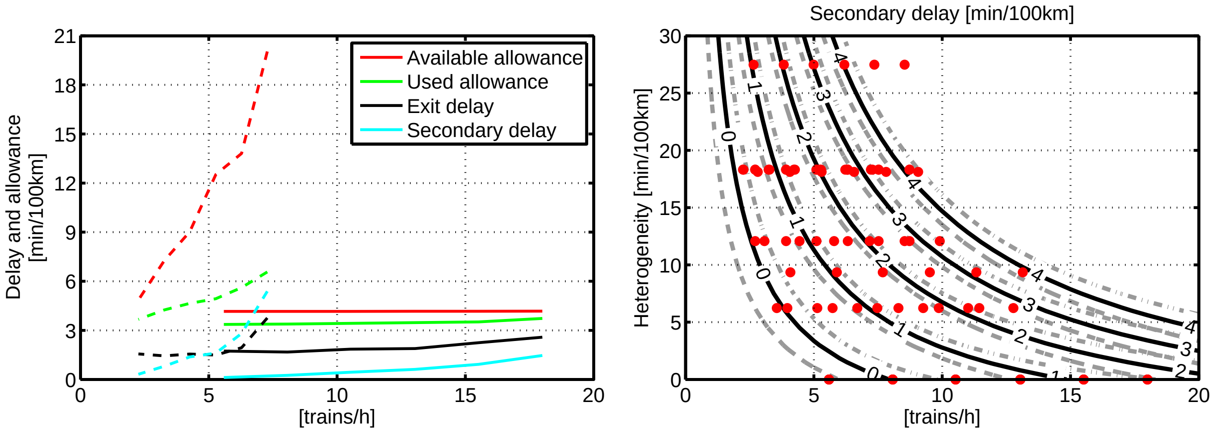


Figure 4.3: Left: Timetable type 1 (homogenous, solid line) and 10 (heterogenous, dashed line), 20 km interstation-distance and high perturbation level.
Right: All timetables, 20 km interstation-distance and high perturbation level.
Contours of fitted surface (black lines) and simulated values (red dots).
Grey contours indicate the 95% prediction intervals of the surface fit.

In paper II a new measure of heterogeneity is introduced, Mean Difference in Free Running time (MDFR). It is calculated as the average of the difference between the running times of the trains in the timetable. The right part of figure 4.3 shows secondary delay as a function of no. trains/h and heterogeneity (MDFR). 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 show 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 homogenous 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 the fact 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 heterogeneity measure that is independent of the number of trains per hour is favoured. It also needs to be stated that the location of the simulated timetables limits the area where the fit is valid. The upper limit is decided by the minimum headway, given the current train types and infrastructure. The lower limit derives from limiting the number of simulations in the experiment.

4.2.2 Primary delay

The impact of primary delays on the creation of secondary delays is shown in figure 4.4, 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 number of lost trains/h, timetables close to maximum capacity are more sensitive to primary delays, which is natural due to smaller buffer times. Homogenous timetables at high traffic densities are sensitive to primary delays due to 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.

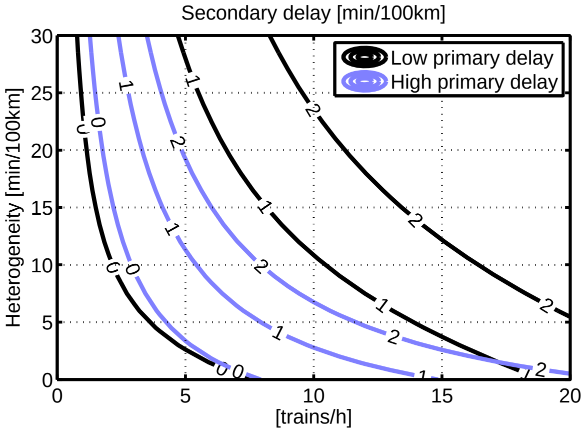


Figure 4.4: Effect of primary delay level on secondary delays.

4.2.3 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 4.5. 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 4.5.

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.

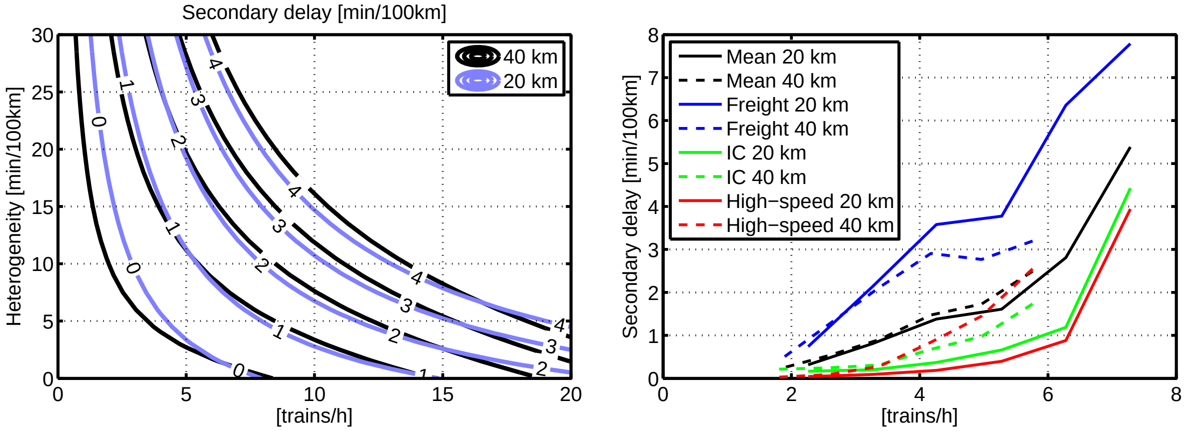


Figure 4.5: Left, effect of inter-station distance on the total secondary delay.
Right, effect of inter-station distance for one type of timetable, separated for the different train types.

4.2.4 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.

Figure 4.6 shows results from timetable type 5, separated for each train type. It consists of freight trains and high-speed trains and is the most heterogeneous of all timetables investigated. The left diagram shows results for the freight trains and the right for the high-speed trains. Several interesting observations are worth commenting:

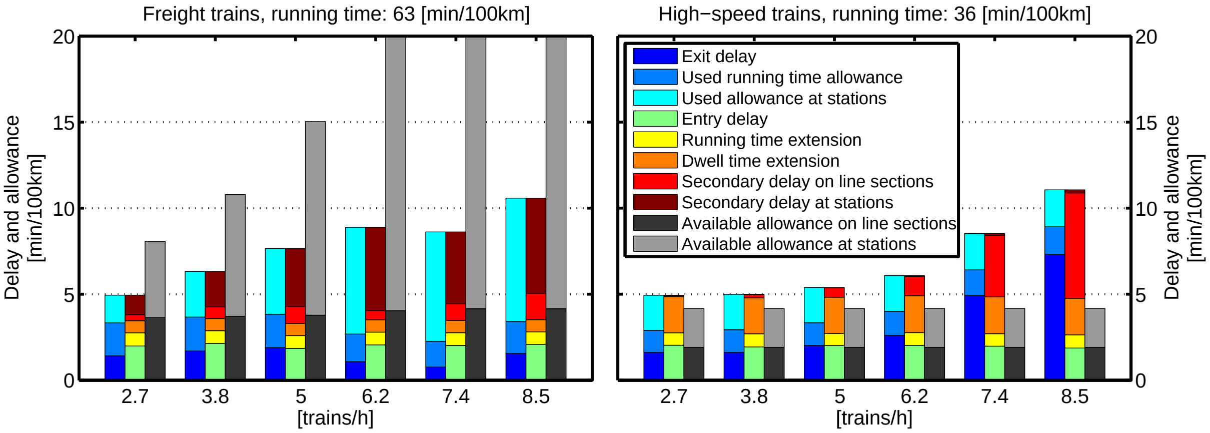


Figure 4.6: Results for individual train types in timetable type 5. Values for the bars ending outside the left figure are (left to right) 37, 39 and 30 minutes.

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 ovartakings 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.

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 4.6.

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 departure 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 4.5 does not show the same timetable as figure 4.2, the behaviour of the involved train types are the same. In figure 4.2, the exit delays remain stable at first, and then start to increase at the same time as secondary delays on line sections increase. Looking at figure 4.5, this behaviour can now be explained by the fact 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.

4.3 Capacity and timetable stability

In order to define 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 how the delays develop as trains run from origin to destination, the positive effect of scheduled delay is included as well as the negative of secondary delays. The capacity limit has been reached if the rate of accumulated delay is too high, i.e. the timetable not stable. 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 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 4.7 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.

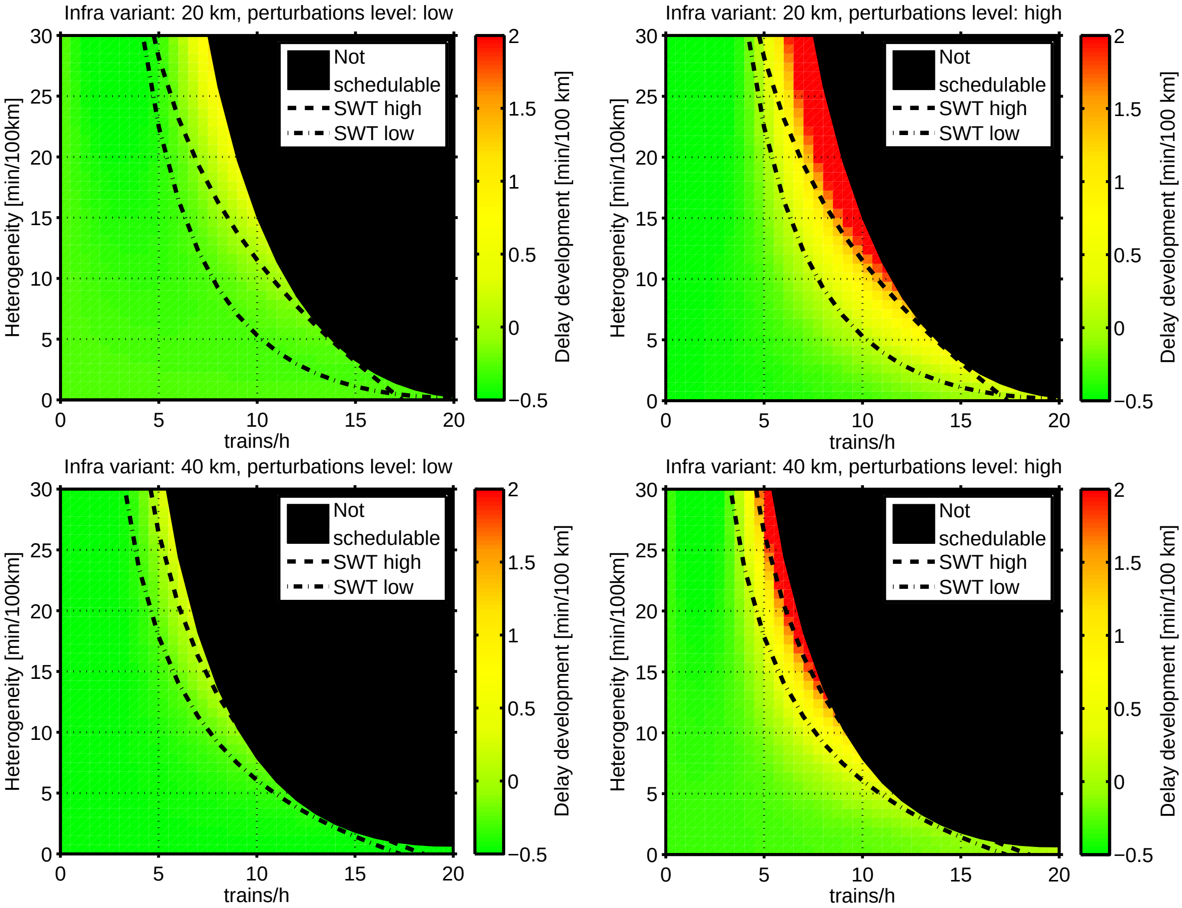


Figure 4.7: Left column, low perturbation level. Right column, high perturbation level. Each row corrsesponds to one infrastructure variant. The color indicates the delay development. Note that the color bar saturates at -.5 and + 2 min/100km, and that lower and higher values are possible. Accepted SWT 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.

Figure 4.7 shows four different scenarios that correspond to the combinations of high and low perturbation levels and two 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 SWT, which is seen as the distance between the black area and the other lines. The extra train slots also suffer from high delays.

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

- For the low primary delays the capacity is limited by the acceptance of SWT. For high level of primary delays, also delays become a limiting factor.

- Inter-station distance has some effect on delay development. The difference between the 20 km and 40 km variants is around 1-2 trains/h if the primary delay level is high.

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 delays need to be included in the analysis of operational delays, also when data from a real network is analysed.

4.4 Capacity and general travel cost

Socioeconomic calculations are frequently used to perform cost benefit analyses of railway infrastructure investments. Costs of the investments are compared to benefits in terms of increased frequency of service and reduced travel times and delays. The general travel cost experienced by a consumer is an important input to the CBA together with travel demand and producer costs (costs for the train operators). General travel cost can be expressed as the weighted sum of ticket price, travel time, delay and waiting time, see equation below. The waiting time is the difference in desired departure time of the consumer and the actual departure of the train service.

$$c = p + \alpha(t + \beta d + \frac{\gamma}{f}) \tag{1}$$

Table 4.3 Parameter values for calculation of general travel cost.

	High-Speed	Intercity/Regional	Freight
α [sek/h]	30 000[3]	10 000[3]	1000[1]
β [-]	3.5[2]	3.5[2]	2

	Passenger trains		Freight
Headway [min]	**0-60**	**60-120**	**-**
γ [-]	0.41[2]	0.22[2]	0

[1] Nelldal and Wajsman (2014)
[2] Trafikverket (2015b).
[3] Value of time is calculated as the average number of passengers multiplied by their average value of time. The average number of passengers on High-speed trains is assumed to be 200 and on Intercity/Regional 100. Average value of time of passengers on High-speed trains is 150 sek/h and on Interciry/Regional 100 sek/h.

In the equation above, c is the general travel cost, p ticket price, t travel time, d delay and f frequency of service. Coefficients α, β and γ are monetary values of travel time, delay and waiting time. The general travel cost together with travel demand is used to calculate the consumer surplus of an investment. The travel time is also used when producer surplus, benefits of the train operators, is calculated. The sum of the consumer and producer surplus defines the total social benefit. When comparing different scenarios in a CBA, the result will depend on what traffic is used when calculating the social benefits. In Eliasson and Börjesson (2014) it is argued that in order to make fair comparison of two infrastructure investments, the traffic (train mix and frequency) should be chosen that maximise the benefit of each scenario. The benefit can be measured in consumer surplus (consumer surplus-maximising), producer surplus (profit maximising) or the sum both (welfare-maximising).

Equation 1 is used to calculate the general travel cost for some of the scenarios simulated in the previous section. Ticket prices are assumed to be constant and can therefore be left out of the calculations. Values of coefficients are summarised in table 4.3. High-speed trains and intercity trains have different value of time, α, because of different number of passengers and different mix of private and business travellers. The value of waiting time, γ, decreases when the frequency of train departures is low. Figure 4.8 shows results for three different traffic mixes (timetable variant 1, 4, and 11 in table 4.2) at varying frequencies on two different infrastructure variants. Looking at the black lines, the same tendency can be observed in all scenarios. General travel cost decreases as the service frequency increase, the main reason being decreased waiting time. However, at some point travel time and delays start to increase faster than the waiting time is decreasing and the general travel cost starts to increase again. Minimum travel cost occurs at different locations for heterogeneous and homogenous timetables. Increased traffic heterogeneity shifts the minimum towards lower traffic densities. The same can be observed when the distance between overtaking stations is increased for heterogeneous timetables.

In order to compare two alternatives, as earlier stated, traffic should be chosen such that the benefit of each alternative is maximised. Consequently the relationship between general travel cost and traffic density is of special importance as it will affect what traffic is chosen for each scenario. With this in mind, it is interesting to observe that the shape of the curves in figure 4.8 changes significantly when train delays are not included in the calculations of general travel cost (red lines). Compared to before, the general travel cost does not start to increase again from a certain point. Therefore, what frequency is considered to be the best will most likely differ depending on if delays are included in the socio economic calculations or not. It should be stated however, that there are many other factors in a socio economic evaluations that affects the optimal frequency, e.g. costs for the train operators, travel demand and that trains have limited capacity (number of seats, maximum length/mass).

It should also be stated that results depend on how the timetables are constructed and that the sequential timetable generator used in this study gives absolute priority to faster trains when timetables are constructed. The consequence is that fast trains receive low increase in SWT but become sensitive to delays as capacity utilisation increases. Faster trains have a higher value of time, table 4.3, and do therefore have higher impact on the results in figure 4.8. This contributes to the large impact of delays in figure 4.8.

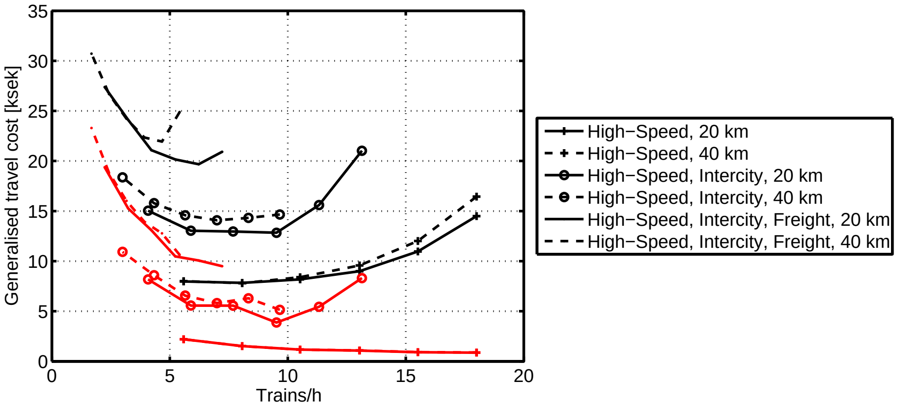


Figure 4.8: Black lines: general travel cost. Red lines: general travel cost excluding delays. The x-axis shows the total number of trains per hour in the timetable (all train train types included).

5 TigerSim capacity model

TigerSim, Timetable Generation and Simulation Model, is a model developed for strategic capacity analysis of double-track railway lines. It is similar to the framework of SAMO presented in section 4.1. It employs a two-step process in which timetables are first generated and then simulated. Compared to SAMO, the most fundamental differences are that timetables are generated by a method similar to synchronous simulation (instead of asynchronous) and that the timetable simulation step does not depend on an external simulation tool (RailSys). Additional differences are listed in table 5.1.

Table 5.1: Differences between SAMO and TigerSim.

	SAMO	TigerSim
Train running time calculation	RailSys	Swedish timetable / manually defined
Infrastructure model	Existing RailSys model / manually defined	BIS / manually defined
Timetable generation	Asynchronous simulation	Synchronous simulation[1]
Timetable simulation	Synchronous (RailSys)	Synchronous (TigerSim)[1]
Infrastructure representation (timetable simulation)	Microscopic	Macroscopic
Timetables per scenario	Single	Multiple

[1] Simulation is not strictly synchronous, see detailed model description in paper V.

The main advantages of the TigerSim model compared to SAMO are:

- Simulation of timetables is faster and model setup is easier.
- The synchronous timetable generation.

SAMO makes it possible to setup and run multiple scenario simulations in RailSys more efficiently, however some steps involving RailSys still require some manual steps, e.g. construction of the infrastructure model, definition of template trains in RailSys, importing timetables and delay data into RailSys and exporting results after simulation in RailSys is finished. Timetable simulations are also significantly faster in TigerSim than in SAMO. The main reason is that TigerSim use a macroscopic infrastructure model and a less complicated and more specialised dispatching functionality. However, because of the macroscopic model, the effect of certain types of infrastructure measures cannot be explicitly studied, e.g. shorter signalling block sections and changed layout at stations etc.

The synchronous approach used in TigerSim for timetable generation makes it possible to prioritise between different types of trains in a more flexible way than is possible in the asynchronous approach used in SAMO. The priority of a train will decide how much SWT it will receive when conflicts with other trains are solved. In SAMO, trains are scheduled in strict order of priority and once a train is scheduled, its train path will not change to accommodate for other trains. The result is that high priority train patterns that are scheduled early in the process will receive no or very low SWT, while low priority patterns have to adapt to all other train patterns already in the timetable, hence receiving very high SWT in congested situations. This is not an unrealistic way to model how timetables are currently constructed in Sweden (Lindfeldt 2009), but it can be inflexible when studying the relationships between SWT, delays and capacity.

5.1 Model description

Capacity analysis in TigerSim is a two-step process, generation and simulation of timetables. In the first step, timetables are generated according to a plan of operation. The plan of operation include information about the different train patterns in the timetable such as origin and destination, frequency, stopping pattern, train type, etc. Based on the plan of operation, a combinatorial approach is used to generate start sequences (combinations of train pattern start times). The method assumes cyclicity and calculates all possible start sequences by systematically shifting train pattern start times relative to each other. The start sequences are input to a scheduling algorithm that generates one timetable for each start sequence. If the number of possible start sequences is too large, a random sample of the possible start sequences can be used. The scheduling algorithm also requires models of the infrastructure and rolling stock, figure 5.1. Generated timetables can be evaluated with respect to some measure of performance, e.g. SWT. If deemed necessary, timetables with poor performance can be discarded before proceeding to the subsequent simulation step. It is then possible to avoid that time is spent on simulation of timetables without interest.

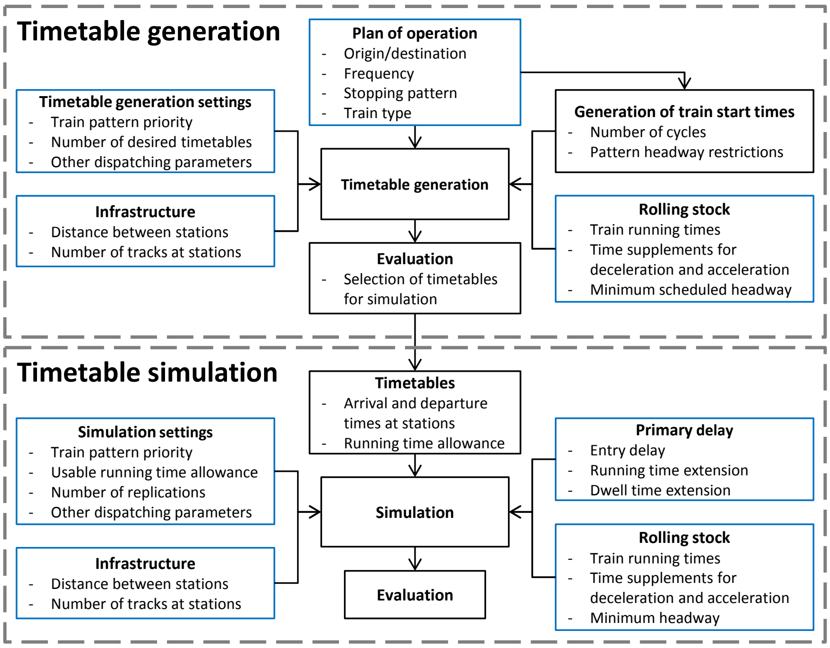

Timetable generation
Plan of operation
- Origin/destination
- Frequency
- Stopping pattern
- Train type
Timetable generation settings
- Train pattern priority
- Number of desired timetables
- Other dispatching parameters
Generation of train start times
- Number of cycles
- Pattern headway restrictions
Timetable generation
Infrastructure
- Distance between stations
- Number of tracks at stations
Rolling stock
- Train running times
- Time supplements for deceleration and acceleration
- Minimum scheduled headway
Evaluation
- Selection of timetables for simulation
Timetable simulation
Timetables
- Arrival and departure times at stations
- Running time allowance
Simulation settings
- Train pattern priority
- Usable running time allowance
- Number of replications
- Other dispatching parameters
Primary delay
- Entry delay
- Running time extension
- Dwell time extension
Simulation
Infrastructure
- Distance between stations
- Number of tracks at stations
Rolling stock
- Train running times
- Time supplements for deceleration and acceleration
- Minimum headway
Evaluation

Figure 5.1: TigerSim model overview. Blue colored boxes indicate model input.

In the timetable simulation step, the scheduling algorithm is reused as a train dispatching functionality. The difference from the generation step is that trains are randomly affected by primary delays and that a feasible timetable is used as reference by the dispatching algorithm when calculating costs for different possible solutions. Primary delays are modelled by stochastic probability distributions and include entry delays, running time extensions and dwell time extensions. Due to the stochastic nature of the primary delays, it may be necessary to simulate each timetable several times, replications, to achieve stable results. Results are evaluated with respect to secondary delays, used allowances, train delay at destination, train punctuality etc.

5.2 Model validation

A necessary step in model development is to validate that TigerSim can simulate timetables accurately. TigerSim is validated in paper V against results from paper III, where SAMO is used to analyse the effect of freight trains skipping scheduled stops. Timetables generated by SAMO in paper III are simulated in TigerSim using the same primary delay distributions, infrastructure and vehicle models, figure 5.2. By comparing results produced by TigerSim with those of RailSys, it is possible to determine the validity TigerSim simulations.

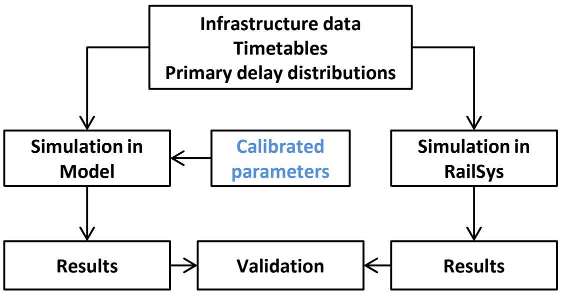


Figure 5.2: Validation workflow.

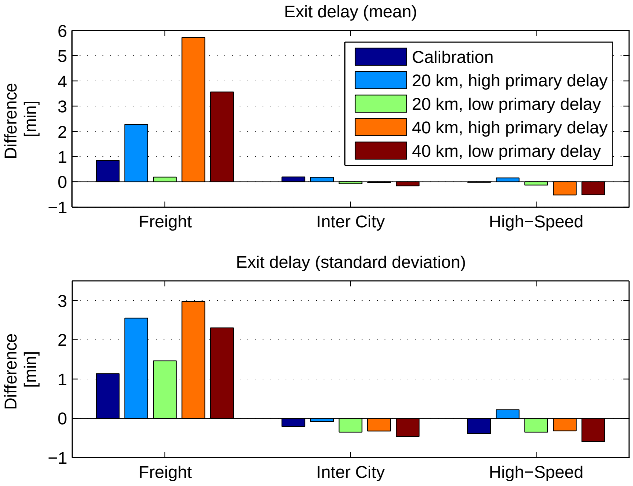


Figure 5.3: Calibration and validation results, mean difference between the model and RailSys. Positive values indicate that delays are overestimated.

Paper V use the same setup as in paper II, thus performed simulations cover a range of different timetables, infrastructure variants and primary delay levels. A subset (timetable 1-6 in table 4.2, all traffic densities, 20 km inter-station distance, high primary delays) of the simulations is used for calibration of parameters in TigerSim. Calibrated parameters are how much of the available running time allowance that can be used to catch up delays, minimum headway between trains and train type priority. Using several timetables with different traffic density and traffic mix makes it possible to calibrate the three parameters individually and in sequence. The remaining simulations are used to validate how the model responds to change in inter-station distance and primary delay levels, figure 5.3. The figure shows the mean difference between the two models for calibration and validation data sets. Results are presented on an aggregate level where each value shown is based on results from several timetables with varying heterogeneity and tragic density. Difference between the two models is greatest for freight trains where the model overestimates the mean and standard deviation of the exit delay.

A detailed comparison between the two models is shown in figure 5.4. Diagrams of this type are powerful and make it possible to understand how both models work and compare them in more detail than is possible if just exit delay is studied. In particular in cases when the two models produce different results, detailed analysis makes possible to understand the reason why. Figure 5.4 shows results from one of the timetable variants with all three types of trains (freight, intercity and high-speed). Traffic load is varied on the x-axis and results are shown separately for the different train types. Corresponding values from the different models are shown next to each other and in the same colour, with RailSys results to the left and TigerSim to the right. The models use the same timetables as input, hence available allowances (grey bars) are the same for both RailSys and TigerSim. Some important observations are listed below.

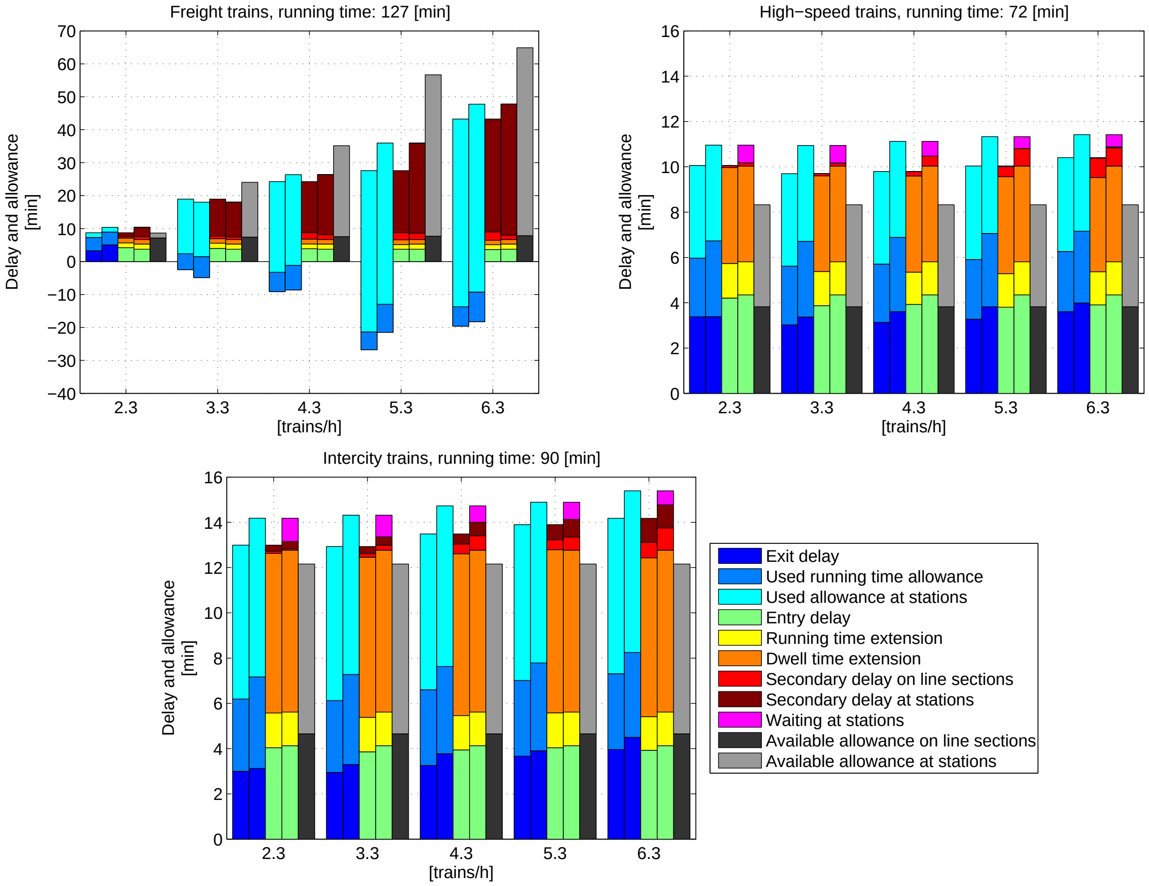


Figure 5.4: Bar graph of simulation results for all five timetables in timetable variant 10, 20 km infrastructure, high primary delays. Results are shown pairwise with RailSys results on the left and TigerSim results on the right bars.

- Same amount of primary delays are applied in both models.

- Calculated secondary delays are similar in both models, both in size and in terms of where they originate, i.e. secondary delays at stations dominate for freight trains, intercity trains suffers from secondary delays on both line sections and at stations and high-speed trains suffer from secondary delays on line sections.

- Freight trains in TigerSim receive more secondary delays than in RailSys.

- More allowance on line sections are used in TigerSim than in RailSys.

Primary delays distributions are input to both models and it should be expected that they are the same. Secondary delay and use of allowance are calculated by the models and are the main reasons why simulations are performed. The models give similar, but some differences do exist. The explanation for these differences is how trains operating ahead of schedule are modelled:

1. In TigerSim, priority is reduced for trains that are ahead of schedule while it remains the same in RailSys.

This is the explanation why freight trains receive higher secondary delays in TigerSim than in RailSys. It also explains why the effect is more evident at higher traffic loads when more freight trains are ahead of schedule.

2. Running time allowance is always available for trains in TigerSim, i.e. it is also available for trains on time or ahead of schedule. In RailSys running time allowance can only be used when a train is late. Consequently passenger trains can never be early in RailSys.

This fact explains why in many cases more running time allowance is used in TigerSim than in RailSys, especially when many trains are early. Another consequence of that running time allowance is always available in TigerSim is that passenger trains can arrive early at scheduled stops, where they have to wait until departure time. In figure 5.4, this time is indicated in magenta.

The differences between the two models discussed above concerns how to model the behaviour of the dispatcher and train driver. It is not possible to say that one of them is more correct. For example, some train drivers may always drive as fast as they can, while others will drive slower if they know they are on time and that there is allowance in the timetable before the next stop. The conclusion of the validation performed in paper V is that the model is valid and can be used for strategic capacity planning.

5.3 Model setup and calibration

Before using the model, it has to be calibrated. The first step of the calibration is to define a plan of operation that reflects the real situation. The plan of operation is then used to generate a set of timetables. In these timetables, departure times at the origin are systematically varied for the different train patterns in the timetable, thus covering a variety of different timetable solutions for each scenario. By generating several timetables, the analysis becomes less dependent on the performance of a single timetable when scheduled and operational delay is estimated, which increases the comparability of the results. In the second step, delay sensitivity of the timetables is analysed using the scheduler in "simulation mode" and the result is evaluated in terms of exit delay. Either the whole set can be used or some criteria can be specified that timetables must meet in order to be used in further analysis and simulation. In this study SWT is used to select appropriate timetables, i.e. if any train pattern has higher SWT than the criteria specified for that pattern, the timetable is discarded and not used for further analysis.

Delay data from real operation is used to compile entry delay distributions (empirical) for the different train patterns in the timetable. Delay data can also be used to estimate distributions for dwell time extensions and running time extensions, but the procedure is not as straight forward. If available, an attractive option can be to use standardised distributions that are known to reflect the current situation. Generated timetables are then simulated with primary delays and results are evaluated. The calibration is an iterative process where model parameters such as train pattern priority (both in the generation and simulation step), minimum headways for different train types are adjusted until results show a good match with real outcome, figure 5.5. The criteria that determine which timetables are accepted for simulation may also have to be calibrated. In figure 5.5, model results are based on the evaluation of 166 timetables while the reference data from real operation is based on a single timetable. This is the explanation why the model shows smoother variations in punctuality development than the reference.

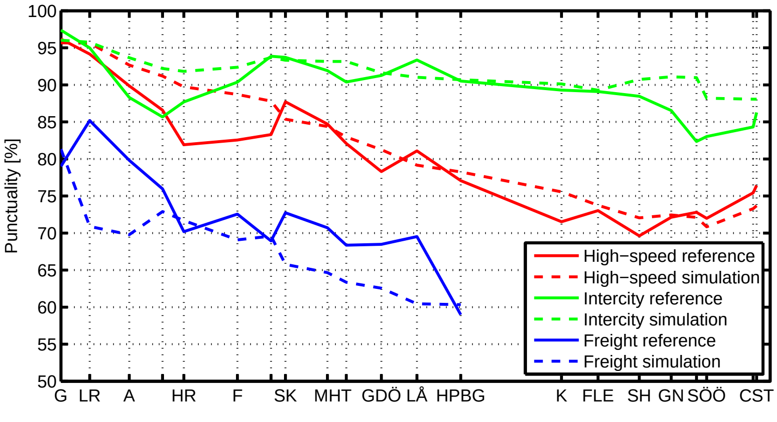


Figure 5.5: Comparison of real outcome (reference) and results from the calibrated model. Patterns running on subsections are included.

The set of timetables chosen for simulation can either be simulated as a group or as individuals. Simulating timetables as a group requires fewer replications (faster), but results are not stochastically stable for individual timetables. It is possible to evaluate results for the timetables as a group, but not for single timetables. Simulating the timetables as individuals takes longer time, but makes it possible to determine the performance of each single timetable and to analyse the spread within the group of timetables. The latter approach is used in the first application below to emphasize the need to base capacity analysis on several timetables. The first approach is used in the second application.

5.4 Application 1: effects of increased high-speed service on a double-track line with mixed traffic

To demonstrate the method, it is applied on a section of the Western main line in Sweden. The Western main line goes from Stockholm to Gothenburg and is one of the busiest double-tracks in Sweden. It is 455 km long and supports a very mixed traffic consisting of high-speed trains (200 km/h), intercity trains of intermediate speed and several slow freight trains and local trains. An example of a real life time distance graph is given in figure 5.6.

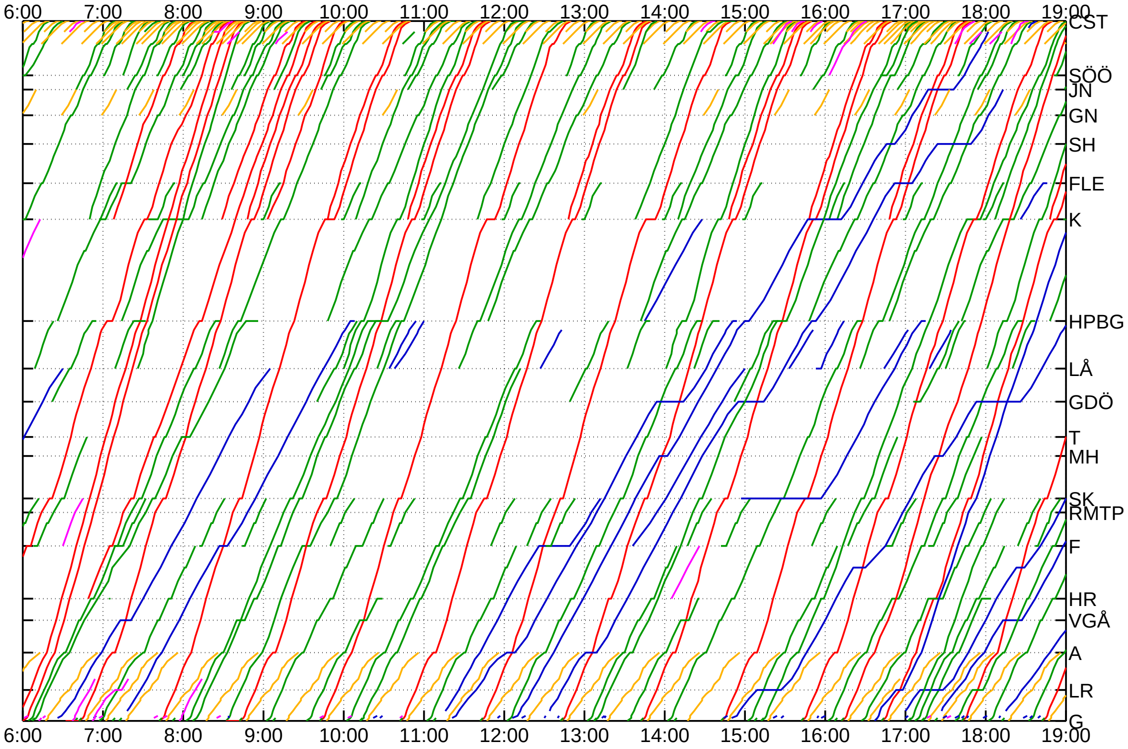


Figure 5.6: Time distance graph for north-bound trains on the Western Main Line, analysed in the case study (T12). Yellow: commuter trains, green: intercity/regional trains, red: high-speed trains, blue: freight trains.

Capacity is analysed by varying the interval of the high-speed service between Gothenburg and Stockholm. The frequency of the high-speed service is varied in six steps from 30 to 180 minutes in steps of 30 minutes. All other traffic patterns are constant and are the same as in the calibration. Table 5.2 summarize some properties of the different scenarios. Due to that many train patterns remain unchanged, the relative difference in total number of operated trains on the line is small, around 7%. However, the increase will have significant impact on SWT and operational delay for several reasons. Firstly, the line is operated close to saturation. Secondly, the varied train pattern of high-speed trains belongs to the fastest train category operated on the line. Adding slow or fast trains to a timetable with a mixed traffic cost more capacity than adding trains with intermediate speed (Gibson et al. 2002). Thirdly, the pattern runs along the entire line from G to CST and can therefore be expected to affect many trains.

Table 5.2: Scenarios in case study For each scenario 10 000 timetables are generated.

Interval of High-speed service [min]	30	60	90	120	150	180
Total number of trains	296	284	280	278	277	276
Number of High-speed trains G-CST	24	12	8	6	5	4
Number of timetables accepted for simulation	11	252	486	410	682	636

The same procedure for selection of timetables for simulation is used in the case study as is used in the calibration stage. Because the same limits on scheduled waiting time are applied in all scenarios, the number of timetables that fulfils them decreases with increasing number of operated trains, table 5.2. In each scenario, 10 000 timetables are generated.

The average SWT for all generated timetables and all scenarios is shown in figure 5.7. The best timetables have between 2-4% of SWT and the worst almost 10%. However, since the figure show average values for all train patterns in the timetable, it is possible that individual train patterns have considerably higher SWT, illustrated in figure 5.10. Even if the total SWT of a timetable is acceptable, it is not likely that a timetable with high SWT for a single train pattern is accepted in reality. It is for this reason necessary to limit the maximum SWT for each train pattern individually to ensure that timetables are representative. Timetables that meet all criteria are shown to the right in figure 5.7. Naturally, timetables with high SWT have been discarded, but it is also possible to conclude that many timetables with low total SWT have been discarded as well since only a few percent of the generated timetables are accepted.

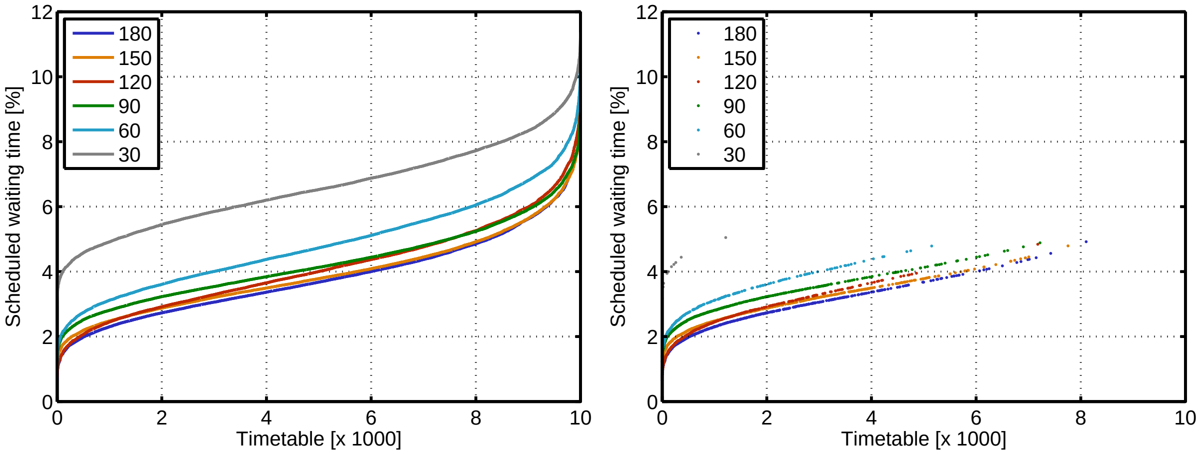


Figure 5.7: Scheduled waiting time of generated timetables for the different scenarios. All timetables (left) and accepted timetables (right). Scheduled waiting time is normalised by train running time and is the average of all evaluated train patterns.

Figure 5.8 shows simulation results for two of the scenarios when high-speed-trains are operated twice and once an hour. Results are separated on different train groups and it is possible to see both SWT and mean delay for each simulated timetable. Black circles indicate mean values for each train group in each scenario. The additional effort taken to simulate each timetable separately makes it possible to determine mean delay of individual timetables. If timetables are simulated as a group, they can only be represented by their common mean value of delay (y-coordinates of circle markers). SWT of individual timetables can always be established as it is a result of the timetable generation step. Just knowing the mean value of SWT and delay may suffice in many analyses. In such cases it is possible to simulate timetables as a group and thereby reduce the time required for simulation considerably. The speed of the analysis can be increased further if no more timetables are generated than is required to get sufficiently many that fulfil the quality criteria and can be used to calculate a stable mean value of the SWT.

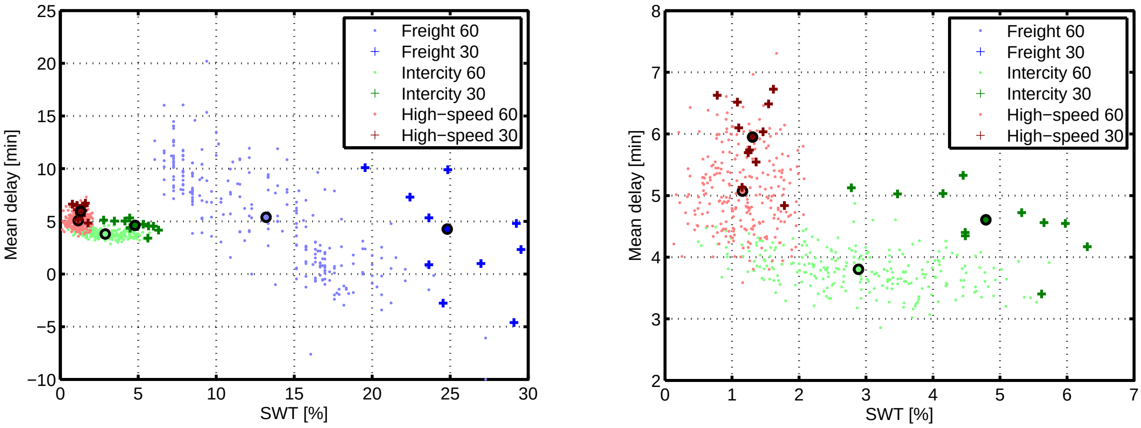


Figure 5.8: Simulation results for different train groups when high-speed trains run every 30 and 60 minutes. Right figure shows same data as the left, but zoomed in on passenger train groups. Circle markers show mean values of each group.

The spread in performance of the timetables showed in figure 5.8 is large relative to the effect of increasing the number of high-speed trains, especially in the case of the high-speed trains where the timetables from the 30 min scenario can be found within the group of timetables from the 60 min scenario. This serves to illustrate the benefit of evaluating several timetables in each scenario and why comparing mean values gives more reliable results. Another interesting aspect is that delays and SWT have a negative correlation, at least for freight trains and intercity trains. This highlights the connection between SWT and delays and is one of the reasons why both need to be considered when capacity is analysed. Otherwise it is possible that the effects of high capacity utilisation are partially hidden in the variable not considered. The connection between SWT and delay is analysed further in section 5.5.

Figure 5.9 is similar to figure 5.8, shows data for all passenger trains together and is presented in the form of boxplots. The boxes show the median (red), 25th percentile and 75th percentile (blue). Whiskers (black) extend to maximum $q_3+1.5\cdot(q_3-q_1)$ and $q_1-1.5\cdot(q_3-q_1)$ where q_1 and q_3 are the 25th and 75th percentiles. Observations outside whiskers are considered outliers. Mean values are indicated with a green line. The boxed makes it easier to compare scenarios and it is possible to conclude that there is considerable overlap between the scenarios both in terms of SWT and delays. The increase in delay is in the same order as the increase of SWT when the number of high-speed trains is increased, which further illustrates the need to consider both SWT and delays in capacity evaluation of double-track lines.

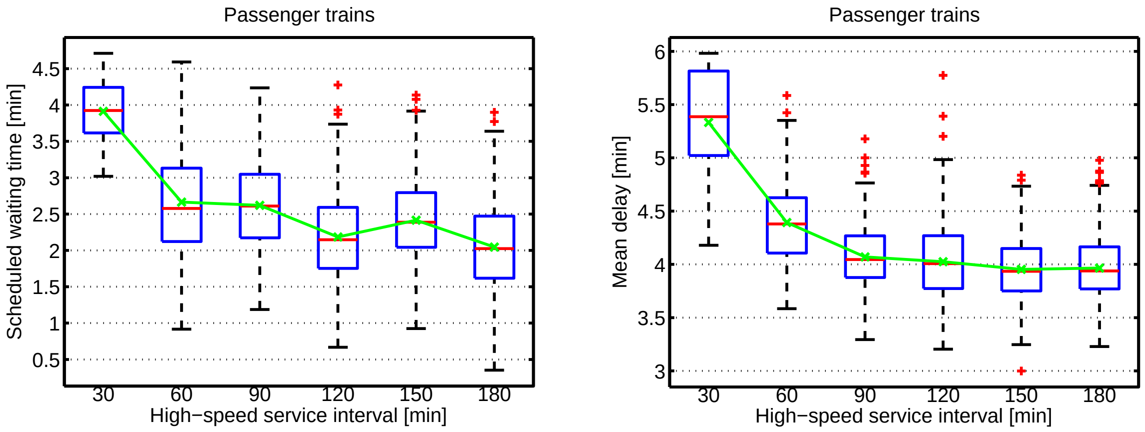


Figure 5.9: Boxplots with scheduled waiting time (left) and delay (right) for passenger trains. Green line indicates mean values.

5.5 Application 2: SWT and delays in double-track operation

Conflicts between trains are solved both by timetable planners when making timetables and by the dispatchers in the operational phase. Basically, the problem is the same and is normally solved by either letting the slower trains stand aside and wait to be overtaken or force the faster train to slow down behind a slower train. Both cases introduce delays for one or both of the trains involved in terms of SWT when timetables are constructed or delays when trains are operated.

There are some differences between the two types of delay however. First, SWT and delays have different value of time in socio economic evaluations. Second, SWT does not necessarily only have to be a bad thing. Even if longer scheduled travel times are associated with costs, SWT can have the positive effect of acting as allowance in the timetable that can be used to recover from delays. It is for this reason an interesting question to ask how trains of different speeds should be prioritized when timetables are constructed in order give more efficient operation. A small increase in SWT for faster trains, might lead to a significant reduction in SWT for slower trains at the same time as delays are reduced for faster trains

The heterogeneous traffic on the Wester main line makes it a good candidate for a study of the relationship between SWT and delays. A subsection of the line, between Gothenburg and Hallsberg, where freight trains are frequent is studied. Timetables are generated using four different setups of time weights used when timetables are constructed, table 5.3. A train pattern with higher time value, relative to the other patterns, is given higher priority and will consequently receive less SWT. Time weights used in simulation is constant in all scenarios.

Table 5.3: Time weights for the different train patterns and different priority settings. The last row indicate the priority used in the simulation.

	Regional	High-speed	Freight
Priority 1	10	100	1
Priority 2	5	25	1
Priority 3	5	15	1
Priority 4	3	9	1
Simulation	10	20	1

Table 5.4: Traffic patterns of through trains in the case study. Numbers indicate trains per hour. Regional trains / high-speed trains / freight trains

Trains/h	**4**	**5**	**6**	**7**
	2/1/1	2/2/1	3/2/1	4/2/1

Table 5.4 summarizes the frequencies of the through trains in the different scenarios. Note that each incremental increase in capacity utilization is achieved by adding either a high-speed train or a regional train, not one of each simultaneously. This is important to keep in mind when analysing the result, as the incremental response is not only explained by the increase in total number of trains per hour, but also by the specific type of train added.

A set of several timetables is generated for each scenario. The real timetable in this case study can quite well be approximate as cyclic, and with a fairly short cycle time, 60-120 min. It takes too big an effort to make timetables for all start sequences. Instead, a random sample of the start sequences is made. The number of samples necessary varies depending on what properties of the timetables are evaluated. In this case, it is the mean value of the SWT and delay, and a sample of 100 is enough to get a satisfactory small standard error of the mean. Figure 5.10 shows examples of how SWT is distributed between different train patterns for the sampled start sequences. SWT for different train patterns are shown in different colour for each timetable, x-axis. SWTs of the train patterns are stacked, i.e. the total height of the bars is the sum of the SWT for all train patterns. Timetables are sorted on total SWT from low (left) to high (right).

The left figure shows results when faster trains are highly prioritised and the right figure when the difference in priority is not as large. Comparing the two figures reveals the trade-off with respect to SWT that exists between train patterns of different speeds. A priority setting that yields low SWT for faster trains (red) give higher SWT for slower trains (blue) and vice versa. The differences between individual timetables in the set highlight the necessity of using several timetables in the analysis. The figures also show that even if two timetables have similar total SWT, it can be distributed differently between the train patterns in the timetable. This motivates why it can be preferable make a selection of appropriate timetables by looking at the individual train patterns instead and not total SWT.

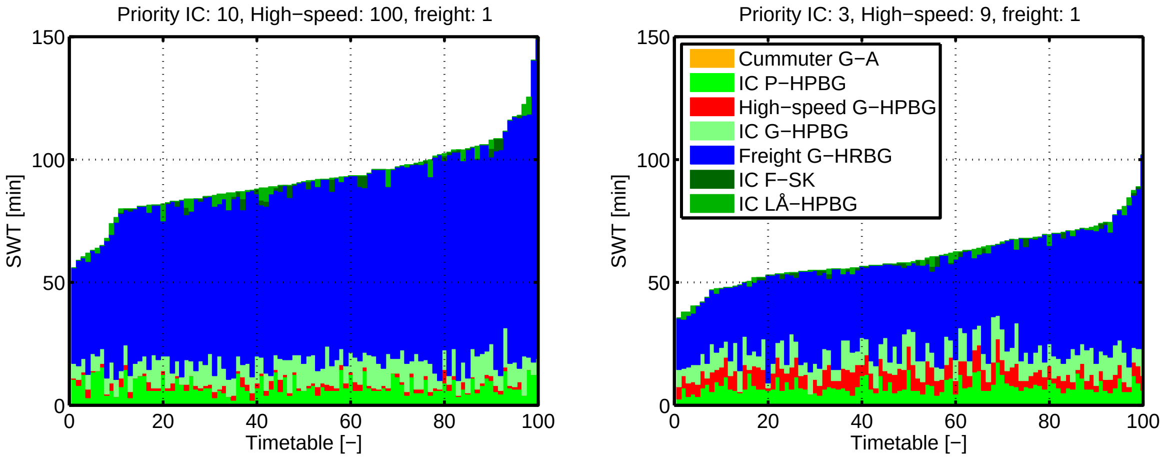


Figure 5.10: Distribution of SWT between different train patterns for the sampled start sequences. Left figure show results using priority setting 1 and right setting 4.

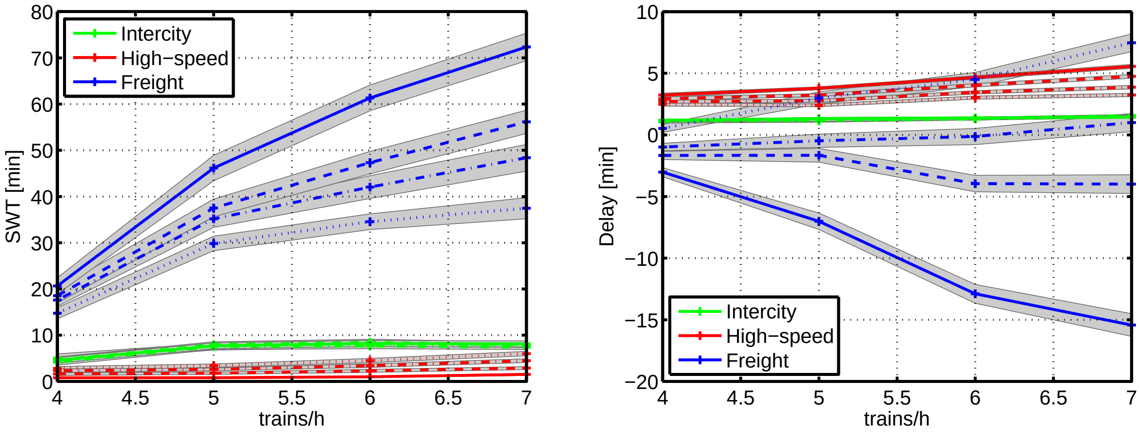


Figure 5.11: Effect of different priority settings and traffic intensity on SWT and average delay. Solid lines: priority setting 1, dashed: 2, dashdot: 3, dotted: 4. Grey area indicate 95% confidence intervals. Negative delays indicate trains arriving ahead of schedule.

Figure 5.11 shows how SWT and delay react when traffic intensity is increased. Results are separated on different train types and priority settings. Traffic intensity is increased from the lowest level by first adding a high-speed train and then a regional train twice in sequence. Adding a train of one train type does mainly affect the SWT of trains of different type, which is natural since trains primarily interact with trains with other speed profiles. Freight trains react more strongly when a high-speed train is added compared to when an intercity train is added. This is due to that high-speed trains have higher speed and priority than intercity trains.

There is a balance with respect to SWT between train patterns of different speeds. If one train pattern is prioritised, other will suffer and get higher SWT. When faster trains are prioritised in priority setting 1, freight trains get a SWT of 72 min, corresponding to an almost 45% increase of the scheduled running time. Using priority setting 4, the corresponding time is reduced to 37 min. The decrease for the freight trains comes at the expense of the high-speed trains that receive 4.5 min longer travel time. However, the increased SWT has the benefit of reducing the average delay of the high-speed trains with 2.5 minute. The regional trains are not affected significantly by priority. This is an effect of that regional trains always have higher priority than freight trains but lower than high-speed trains. Most of the changes in priority favours the regional trains relative to one train type at the same time as it becomes less beneficial compared to another.

By using the value of time from table 4.2 (α and β) makes it possible to add SWT and delays for the different train types together and compare the different priority settings from a socio economic perspective. Figure 5.12 shows the delay cost and it is possible to read out how much capacity can be gained by choosing the right priority scheme and how it varies with traffic intensity. Priority setting 1 is the best setting if only SWT is considered while setting 4 is best when delays are also included. The difference between the settings is smaller when capacity utilisation is low.

The case study shows that it might be motivated in some cases to reduce the priority of the faster trains when timetables are made. This will extend their travel time, but they will benefit from reduced delays as well as allowing other slower trains to get shorter travel times.

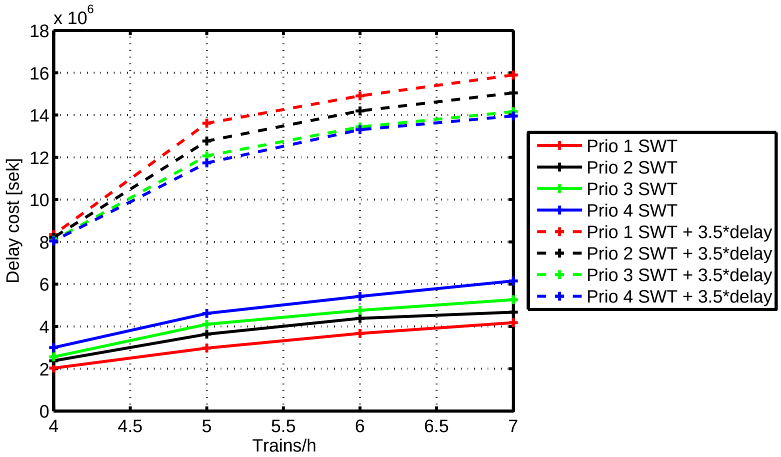


Figure 5.12: Total delay costs for different priority settings and traffic intensities. Colours indicate priority setting and the line style whether or not delays are included in the calculation of delay cost.

5.6 Conclusions

This chapter introduces TigerSim, a new model for capacity evaluation of double-track railway lines. Compared to the model introduced in the previous chapter, TigerSim does not make use of RailSys for timetable simulation. Instead, timetables are simulated within the model using the same macroscopic infrastructure model as used when timetables are generated. As a result, the time needed for setting up the model and performing the simulations is significantly reduced. The increased speed makes it possible to analyse many scenarios where each scenario is represented by multiple timetables. Evaluating many timetables in each scenario makes results less dependent on the performance of single timetables and thereby more reliable in situation when the timetable is unknown.

In each scenario multiple timetables are generated and simulated according to a plan of operation. The plan of operation defines traffic in terms of different train patterns and their associated properties, such as frequency, origin/destination, train type, stopping pattern, primary delay distributions, priorities etc. Results can either be evaluated with respect to single timetables or to the plan of operation. A smaller effort is required if results are evaluated with respect to the plan of operation as timetables do not need to be simulated with as many replications as is necessary if each timetable is evaluated individually. It is then possible to determine how the plan of operation affects SWT and delays on average without being dependent on the performance of a single timetable. An example of application is future travel demand estimation, where the correlation between train service frequency and travel time and delays are of interest.

Simulating all timetables to stability makes it possible analyse the spread in performance of the timetables in each scenario, illustrated in figure 5.13 to the left. It is then possible to do a more detailed analysis of the importance of the timetable and better comparisons between alternatives. If for example two infrastructure alternatives are compared, it is possible to compare not only the mean value of the two sets of timetables, but also the spread. An infrastructure with small spread may indicate that it is more flexible with respect the timetable compared to another with a larger spread, assuming that mean values are similar. Timetable flexibility is an important factor since it may not always be possible in reality to select a timetable with good performance, e.g. because of connections and transfers to other railway lines. The spread of two alternatives can also be used to determine if the difference between them is significant.

TigerSim does not explicitly calculate capacity in terms of how many train can be operated per hour. However, it can be used to determine if a traffic scenario will fulfil specified requirements on quality of service in terms of increased travel time (SWT) and delays. Capacity can be calculated by systematically increasing traffic load until quality of service can no longer be met. Data from real operation can be used as input to and calibration of the model to make it reflect current or expected conditions including primary delays and policies for timetable construction and train operation.

Output from TigerSim can be used as input to socio economic evaluations of different traffic scenarios or infrastructure investments. Figure 5.13 (right) shows an illustrative example, where general travel cost is calculated applying the procedure described in section 4.3.1 on the results in section 5.4. The upper curve shows results based on both SWT and delays and the lower curve when only SWT is considered. The difference from the previous calculations is that each individual case is based on results from several timetables. This improves the possibility of comparing different alternatives in a better way as it is possible to reduce the impact of the performance of individual timetables, which might otherwise lead to false conclusions. In figure 5.13 the median is indicated by the red line and the variation is indicated by percentiles, 10 % steps, in shades of blue colour. In the upper curve, minimum general travel cost is achieved when high-speed trains are operating at 60 min interval. However, the variation in timetable performance is considerable relative to the effect of varying the frequency of the high-speed service. Similar to the results in section 4.3.1, no minimum is reached when delays are not included.

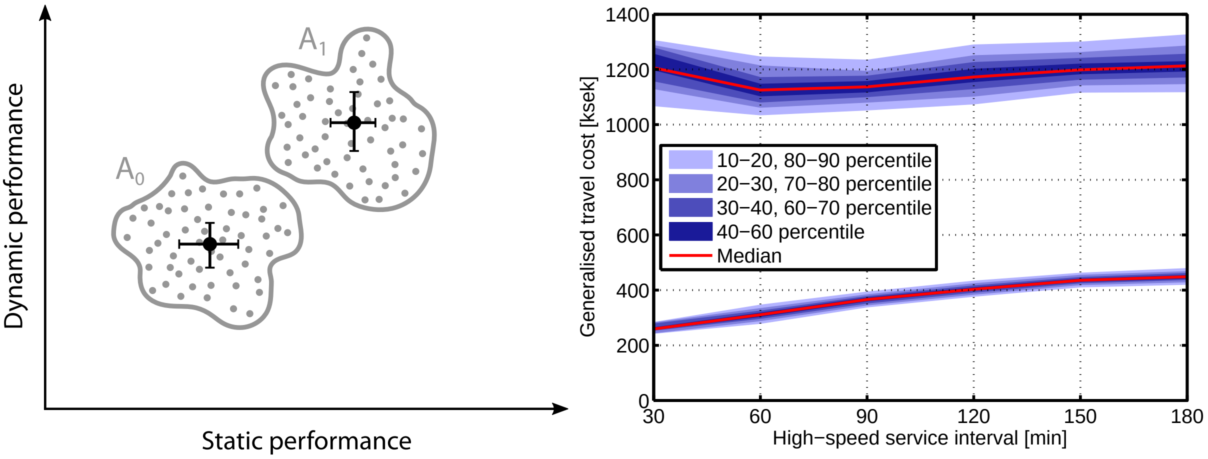


Figure 5.13: Left: schematic picture showing two groups of timetables with different static and dynamic performance. Right: general travel cost for the scenarios in section 5.4. : Top line show travel time + delay +waiting time. The lower line show results when delays are not included in the calculations.

6 Contribution of thesis

6.1 Methodological framework

In paper II a method is developed that expands the potential of using microscopic simulation to perform more general and less timetable dependent analyses. 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. 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. The interface to the micro simulation tool RailSys created in paper II together with a timetable generation algorithm makes it possible to simulate hundreds of different timetables, 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.

Train traffic heterogeneity is a key factor for capacity consumption on double-track railway lines. In this thesis, the effect of heterogeneity in combination with traffic load is studied in detail. Extensive simulation experiments are used to determine how it affects train operation with respect to SWT, secondary delays and use of allowance under different conditions. In paper IV special interest is devoted to how different measures of heterogeneity can be used to explain secondary delays. The paper is one of few studies that compare several different measures of heterogeneity found in literature and analyse how they can be used to explain secondary delays under a wide range of conditions. Several new measures for heterogeneity are proposed in this thesis. An advantage of one of the new measures is that it can be used to predict how trains of different speed profiles are affected by secondary delays.

The power of the two methods for capacity analysis developed in this thesis lies in the combination of timetable scheduling and timetable simulation. The method developed in paper V-VI, TigerSim, generates and simulates several timetables per scenario which makes it possible to draw more general conclusions in cases where it is not desirable that results are dependent on the performance of a specific timetable. The model can be used to calculate capacity of double-track railway lines taking several important aspects of railway operation into consideration:

- Infrastructure (double-track)
- Plan of operation
- Policies of timetable construction (e.g. train priority, buffer times, allowances)
- Policies of train operation (e.g. train priority)
- Vehicle performance
- Primary delays
- Required quality of service (Scheduled waiting time and train delay)

Another aspect of contribution is the analysis of simulation results. In several of the papers, simulation results are evaluated with respect to not only train delays, but also to secondary delays and use of allowance along the route. This data, together with primary delays and available allowance, give a more comprehensive understanding of the results as well as insights into how the railway system works in different situations.

6.2 Results of applied models

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. The work of paper I was followed up by a new study in 2013 (Lindfeldt, 2014). Results have been used as input to other studies and to improve data quality of some of the Swedish Transport Administration's databases.

Paper III presents a method that makes it possible to model freight train operations more realistically. Results indicate that passenger trains are not affected negatively when freight trains are allowed to depart ahead of schedule. Results are consistent with a Swedish field study performed in 2009 (Banverket, 2009). The new methodology is of use when freight train operation is the focus of the study.

Capacity of a double-track line is reduced considerably if traffic is traffic is heterogeneous, i.e. train patterns operate at different average speeds. As traffic intensity increases, more allowance has to be allocated to the trains in order to make a conflict free timetable. With today's operational principles in Sweden, faster passenger trains are often prioritised before slower freight trains. The result is that travel time is extended for freight trains and that faster trains become sensitive to delays. In practise this limits capacity to 5-7 trains/h on double-track lines with heterogeneous traffic, compared to a theoretical capacity of 18-20 trains/h if traffic is completely homogenous. Other conclusions about double-track railway operation are:

- If primary delay levels are low, capacity is limited by how much SWT is considered to be acceptable. If delay levels are high, also timetable stability (delay growth) becomes a limiting factor.
- In heterogeneous traffic, it can be beneficial to allocate more allowance to the fastest trains. The high speed trains will receive longer scheduled travel time, but in return their delay will decrease at the same time as scheduled travel times for slower trains are reduced.
- The distance between stations where trains can pass each other (40/20km) affects the total amount of secondary delays only marginally. However, secondary delays for high-speed trains will increase faster with higher frequency if the inter-station distance is longer. At the same time freight trains receive less secondary delays.
- The distance between stations where trains can pass each other (40/20km) affects the generalised travel cost of heterogeneous traffic. Lower cost can be achieved with shorter inter-station distance and the minimum cost occurs at higher traffic intensity.
- It is possible to schedule more trains/h with shorter inter-station distances. This effect is more evident for heterogeneous timetables (approx. 50% more in the most heterogeneous case). However, the extra train slots have high levels of SWT and are very delay sensitive.

In Sweden commercial long distance rail traffic is deregulated. In 2015 a competing service with the Swedish state railways high-speed service will start between Stockholm and Gothenburg resulting in that the frequency of the high-speed trains will be almost doubled from one to two trains/h during parts of the day.

The TigerSim model has been used to analyse the effect of changing the frequency of the high-speed service between Stockholm and Gothenburg. The interval is varied in 30 min steps from 30 to 180 minutes while keeping all other traffic constant. Increasing service frequency from one train/h to two trains/h during the whole day has the following effects:

- The flexibility to construct good timetables is reduced considerably. The number of timetables with acceptable performance found by TigerSim decreased from 252 to 11 (out of 10 000).
- Scheduled waiting time increase from 2 to 3% for passenger trains and from 13 to 25% for freight trains operated during the daytime.
- Mean delay for passenger trains increase with approximately 20%.

Calculations of generalised travel cost taking travel time and waiting time into consideration show that the lowest cost is achieved when the high-speed trains are operated at 30 minutes interval. However, if train delays are also included in the calculation of generalised travel cost, the minimum cost will be at 60 minutes interval. Even if there are more factors that should be taken into account in a socio-economic calculation, e.g. operation costs, limited seat capacity of the trains and the possibility to lower the price by competition, results indicate the importance of taking train delays into account in the analysis.

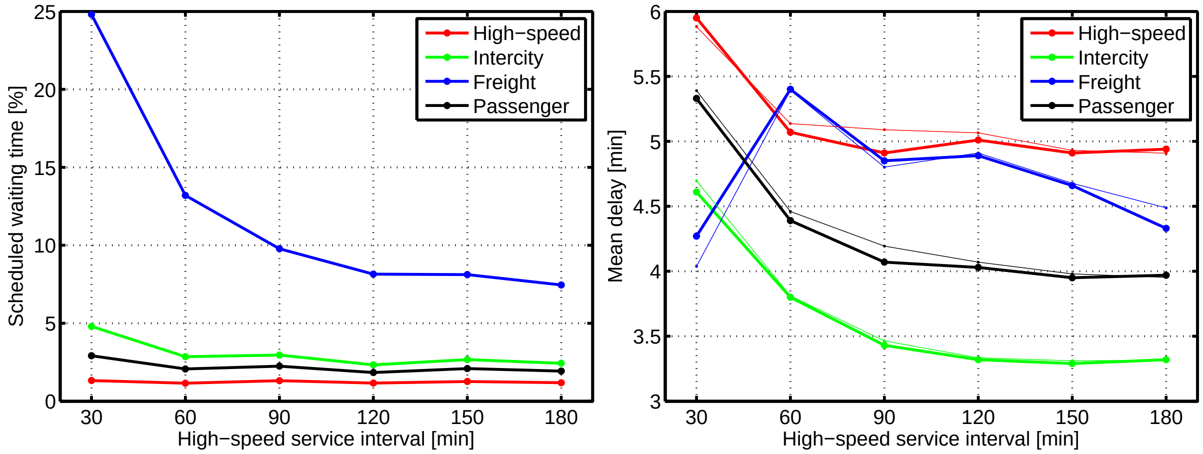


Figure 6.1: SWT and mean delay for different train types when high-speed service interval between Gothenburg and Stockholm is varied.

7 Future work

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. 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 etc.

The method developed in paper II that allows timetables and perturbations to be directly imported into RailSys opens up several new opportunities. For example can the impact of primary delays be studied in detail. In paper II and IV all types of primary delays are varied together, which makes it impossible to distinguish between 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.

The TigerSim model can be further developed in several aspects. One improvement that would increase the realism of the model, and that would not be too difficult to implement, is to include crossing traffic at junctions. Crossing traffic affects traffic running on the main line and cause secondary delays. Another possibility that would reduce the time necessary for setting up the model is to implement an automatic procedure for calibration of model parameters. A suitable optimisation method can be applied to calibrate e.g. train pattern priorities, minimum headways and primary delay distributions with the objective of minimising the difference between model output and data from real operation. A third potential improvement is to consider connections between trains in a more sophisticated way. It is then possible to create more realistic timetables that can be evaluated with respect to travel time and delays at a passenger level instead of at train level.

TigerSim can be complemented with better socioeconomic evaluations of results. SWT, delays and service of frequency are currently only used to calculate a generalised travel cost. However, in order for results to be useful in e.g. a CBA of an infrastructure investment, more sophisticated socioeconomic calculations are needed that include e.g. producer costs and travel demand modelling etc. In paper VI SWT is used as measure of performance to select timetables that are simulated with delays. The selection can be improved if a suitable measure of timetable robustness is also included in the evaluation, e.g. the SAHR (Vromans et al. 2006) or the RCP (Andersson et al. 2013). Timetables with poor robustness can then be disqualified before simulation if they are deemed unrealistic.

TigerSim is developed for analysis of double-track railway lines. The usefulness of the model would increase if it can be made to also handle opposing traffic on single-track lines. However, this would require a lot of work as a completely new scheduling/dispatching functionality needs to be implemented. A better approach might then be to develop a complete new model using existing heuristics or optimisation based algorithms for timetable generation in combination with existing advanced micro simulation tools. The model can then be made general enough to handle analysis of railway networks. Generated timetables do not need to be optimal, only good enough to be realistic representations future timetables. If timetables are simulated as a group, it should be possible to make the model fast enough to handle multiple timetables per scenario in the same way TigerSim does.

8 References

Abril, M., Barber, F., Ingolotti, L., Salido M.A., Tormos, P., Lova, A., 2008. “An assessment of railway capacity”. Transportation Research Part E, 44, 774-806.

Andersson, E.V., Peterson, A., Törnquist, Krasemann, J., 2013 “Quantifying railway timetable robustness in critical points”, Journal of Rail Transport Planning & Management, vol.3, pp. 95-110.

Andersson, E.V., Peterson, A., Törnquist, Krasemann, J., 2015 “Improved Railway Timetable Robustness for Reduced Traffic Delays – a MILP approach”, In proceedings of 6:th International Seminar on Railway Operations Modelling and Analysis, Tokyo, Japan.

Banverket, 2009. “Testvecka: släpp inte ut tidiga tåg 2-6 mars (Test week: no clearing of early of early trains March 2-6). Swedish Rail Administration, Internal report. (in Swedish).

Barter, W.A.M. 2008, "ERTMS Level2: effect on capacity compared with "best practice" conventional signalling", Proceedings of the 11th International Conference on Computers in Railways, eds. J. Allan, E. Arias, C.A. Brebbia, et al, WIT Press, Great Britain, pp. 213.

Büker, T., Seybold, B., 2012 “Stochastic modelling of delay propagation in large networks”, Journal of Rail Transport Planning & Management 2 (2012), 34–50.

Burdett, R.L., Kozan, E., 2006. “Techniques for absolute capacity determination in railways”, Transportation Research part B, 616-632.

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., Wagenaar, J., 2014 “An overview of recovery models and algorithms for real-time railway rescheduling”, Transportation Research Part B 63, 15–37.

Cerreto, F., 2015. “Micro-simulation based analysis of railway lines robustness”, In proceedings of 6:th International Seminar on Railway Operations Modelling and Analysis, Tokyo, Japan.

Cui, Y., Martin, U., 2011. “Multi-scale simulation in railway planning and operation”, Promet – Traffic&Transportation, Vol. 23, 2011, No. 6, 511-517

D’Ariano, A., Pranzo, M., 2009. “An advanced real-time train dispatching system for minimizing the propagation of delays in a dispatching area under severe disturbances” Networks Spatial Econ. 9 (1), 63–84.

de Fabris, S., Longo, G., Medeossi, G., Pesenti, R., 2014. ” Automatic generation of railway timetables based on a mesoscopic infrastructure model” Journal of Rail Transport Planning & Management 4 (2014) 2–13.

Eliasson, J., Börjesson, M., 2014. “On timetable assumptions in railway investment appraisal”. Transport policy, 36, 118-126.

Fischetti, M., Salvagnin, D., Zanette, A., 2009. “Fast approaches to improve the robustness of a railway timetable”, Transportation Science, 43, 321-335.

Gibson, S., Cooper, G., and Ball, B. (2002). “Developments in transport policy: The evolution of capacity charges on the uk rail network”, Journal of Transport Economics and Policy 36, 341-354.

Gille, A., Klemenz, M., Siefer, T., 2008. “Applying multiscaling analysis to detect capacity resources in railway networks”, In: Allan, J., Arias, E., Brebbia, C.A., Goodman, C., Rumsey, A.F., Sciutto, G., Tomii, N. (eds.), Computers in Railways XI, WIT Press, Southampton, UK.

Gorman, M., (2009), “Statistical estimation of railroad congestion delay”, Transportation Research Part E.

Goverde, R.M.P., (2007), “Railway timetable stability analysis using max-plus system theory”, Transportation Research Part B: Methodological 41.

Goverde, R.M.P., Corman, F., D’Ariano A., 2013. “Railway line capacity consumption of different railway signalling systems under scheduled and disturbed conditions” Journal of Rail Transport Planning & Management, 78–94.

Goverde, R.M.P., Hansen , I.A., 2013. “Performance indicators for railway timetables” In proceedings of International Conference on Intelligent Rail Transportation (ICRIT), Beijing, China.

Grimm, M., Wahlborg, M. (2008), Kapacitetsutnyttjande och kapacitetsbegränsningar 2007/2008, Banverket Leverans, rapport.

Harrod, S., 2009. ”Capacity factors of a mixed speed railway network”, Transportation Research Part E, 45, 2009, 830-841.

Harrod, S., 2012. “A tutorial on fundamental model structures for railway timetable optimization”, Surveys in Operations Research and Management Science, 17, 2012, 85-96.

Huisman, T., Boucherie, R., (2001), “Running times on railway sections with heterogeneous train traffic”, Transportation Research Part B 35, 2001, 271-292

Huisman, T., Boucherie, R.J., van Dijk, N. M., 2002. “A solvable queueing network model for railway networks and its validation and applications for the Netherlands” European Journal of Operational Research, 30–51.

Kaas, A.H., 1998. “Development and practical use of a capacity model for railway networks”, Proceedings of the Conference on Structural Integrity and Passenger Safety, ed. C.A. Brebbia, WITpress, Great Britain.

Khoshniyat, F., Peterson, A., 2015. “Robustness Improvements in a Train Timetable with Travel Time Dependent Minimum Headways”, In proceedings of 6:th International Seminar on Railway Operations Modelling and Analysis, Tokyo, Japan.

Kroon, L., Maróti, G., Helmrich, M. R., Vromans, M., & Dekker, R. (2008). Stochastic improvement of cyclic railway timetables. Transportation Research Part B: Methodological, 42(6), 553-570.

Krueger, H., 1999. Parametric modelling in rail capacity planning. Proceedings of Winter Simulation Conference, Phoenix, AZ.

Lai, Y. C., Liu, Y. H., Lin, Y. J., 2013. ”Development of Base Train Equivalents for Headway-Based Analytical Railway Capacity Analysis” In proceedings of 5:th International Seminar on Railway Operations Modelling and Analysis, Copenhagen, Denmark.

Landex, A. (2008), “Methods to estimate railway capacity and passenger delays”, PhD Thesis, Technical University of Denmark.

Lindfeldt, A. (2008) Kapacitet på linjen, memorandum.

Lindfeldt, A., 2009. Kapacitetsanalys av järnvägsnätet i Sverige, delrapport 2. Capacity analysis of the Swedish rail network, part 2. In Swedish.

Lindfeldt, A., 2014. Kapacitetsutnyttjande i det svenska järnvägsnätet. Uppdatering och analys av utvecklingen 2008-2012. Capacity utilisation of the Swedish railway network. Update and analysis of the development 2008-2012.

Lindfeldt, O. (2010), “Impacts of infrastructure, timetable and perturbations in operation of double track railway lines with mixed traffic”, Proceedings of 12th World Conference on Transportation Research, Lisbon.

Lindfeldt, O., 2009. “Analysis of capacity on double-track railway lines”. Transport Planning and Technology.

Lindner, T. (2011), “Applicability of the analytical UIC Code 406 compression method for evaluating line and station capacity”, Journal of Rail Transport and Planning & Management, p 49-57.

Magnarini, M. (2010), “Evaluation of ETCS on railway capacity in congested areas – A case study within the network of Stockholm”, Master Thesis, KTH

Mattsson L-G., (2007) Railway Capacity and Train Delay Relationships. Critical Infrastructure: Reliability and Vulnerability, Springer-Verlag, 129-150.

Medeossi, G., Longo, G., de Fabris, S., 2011. “A method for using stochastic blocking times to improve timetable planning”, Journal of Rail Transport Planning & Management, vol 1, pp 1-13.

Meng, L., et al., 2013. “Assessing the absolute traffic carrying capacity: a train timetabling approach”, In proceedings of 5:th International Seminar on Railway Operations Modelling and Analysis, Copenhagen, Denmark.

Murali, P., Dessouky, M., Ordonez, F., Palmer, K. (2010) “A Delay Estimation Technique for Single and Double-track Railroads”, Transportation Research Part E: Logistics and Transportation Review, Volume 46, Issue 4, July 2010, Pages 483–495

Nelldal, B-L., Lindfeldt, O., Sipilä, H., Wolfmaier, J. (2008) "Förbättrad punktlighet på X2000 – analys med hjälp av simulering, (Improved punctuality for X2000 – a simulation approach). KTH. (In Swedish)

Nelldal, B-L., Lindfeldt, A., Lindfeldt, O. (2009) ”Kapacitetsanalys av järnvägsnätet I Sverige, delrapport 1”, KTH, Technical report (In Swedish)

Nelldal, B-L., Wajsman, J. (2014) “Utvecklingen av rangerbangårdarna i Sverige –Hittillsvarande utveckling och samhällsekonomiska kalkyler för rangerbangårdar samt prognoser för järnvägens produkter”, KTH, Technical report (In Swedish)

Nock, O.S. 1980, Railway Signalling, A & C Black.

Pachl, J., (2007) “Avoiding Deadlocks in Synchronous Railway Simulations”, In proceedings of 2nd International Seminar on Railway Operations Modelling and Analysis, Hannover, Germany.

Peterson, A., 2012. “Towards a robust traffic timetable for the Swedish southern mainline”, In: WIT Transactions on the Built Environment, vol. 127, pp. 473-484.

Pouryousef, H., Lautala, P., White, T., 2015. “Railroad capacity tools and methodologies in the U.S. and Europe”, Journal of Modern Transportation, vol 23, issue 1, pp30-42.

Radtke, A., Hauptman, D., 2004. Automated planning of timetables in large railway neworks using microscopic data bases and railway simulation techniques. In Computers in Railways IX, WIT press.

Rudolph, R., (2003), “Allowances and margins in railway scheduling” Proceedings of the World Congress on Railway Research, pp 230–238. Edinburgh, Scotland.

Salido, M., Barber, F., Ingolotti, L., 2012. Robustness for a single railway line: Analytical and simulation methods. Expert Systems with Applications 39, 13305-13327.

Sameni MK, Dingler M, Preston JM, Barkan CPL (2011) Profit-generating capacity for a freight railroad. In: TRB 90th Annual Meeting, TRB, Washington, DC

Schwanhäußer, W (1974) Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn, Dissertation, Veröffentlichungen des Verkehrswissenschaftlichen Institutes der RWTH Aachen, Heft 20

Siefer, T.,2008. Simulation, in: Hansen, I., Pachl, J., (Eds.), Railway Timetable & Traffic. Eurailpress. chapter 9, pp. 155-169.

Sipilä, H., 2012. Simulation of rail traffic: applications with timetable construction and delay modelling. Licentiate thesis, KTH Royal Institute of Technology, Stocholm, Sweden.

Sipilä, H., 2014. “Evaluation of single track timetables using simulation”, in Proceedings of the 2014 Joint Rail Conference.

Sipilä, H., 2015. “A simulation based framework for evaluating effects of infrastructure improvements on scheduled and operational delays”, In proceedings of 6:th International Seminar on Railway Operations Modelling and Analysis, Tokyo, Japan.

Sogin, S. L., Barkan, C. P. L., Saat, M. 2011, “Simulating the effects of higher speed passenger trains in single track freight networks”, Proceedings of the 2011 Winter Simulation Conference

Sogin, S. L., Lai, Y.-C., Dick, C. T., Barkan, C. P. L. “Analyzing the incremental transition from single to double track railway lines”, In proceedings of 5:th International Seminar on Railway Operations Modelling and Analysis, Copenhagen, Denmark.

Trafikverket (2015a), Järnvägens kapacitet 2014, Underlag till årsredovisningen, teknisk rapport (Swedish Transport Administration 2015, Railway capacity 2014, Documentation for Annual Report Technical report) In Swedish

Trafikverket (2015b), Samhällsekonomiska principer och kalkylvärden för transportsektorn: ASEK 5.2. Report in Swedish.

Törnquist, J., Persson, J.A., 2007. “N-tracked railway traffic re-scheduling during disturbances”, Transportation Research Part B: Methodological, vol. 41, pp. 342-362.

UIC Code 405, 1996. Links between railway infrastructure capacity and the quality of operations, International Union of Railways.

UIC Code 406 1st edition, Capacity, International Union of Railways, 2004.

UIC Code 406 2nd edition, Capacity, International Union of Railways, 2013.

UNECE, 2014. Number of railway passengers by country, passengers and time. United Nations Economic Commission for Europe, statistical database.

Vromans M., Dekker R., Kroon L., (2006) Reliability and heterogeneity of railway services. European Journal of Operational Research 17, p647-665

Vromans, M. (2005), “Reliability of Railway Systems”, PhD Thesis, Erasmus University Rotterdam

White, T., (2005), “Alternatives for Railroad Traffic Simulation Analysis” Transportation Research Record: Journal of Transportation Research Board, 1916, pp 34-41.

Wittrup Jensen, L., Landex A., Anker Nielsen, O., 2015. “Assessment of Stochastic Capacity Consumption in Railway Networks” In proceedings of 6:th International Seminar on Railway Operations Modelling and Analysis, Tokyo, Japan.

Yuan, J., Hansen, I.A., (2007), “Optimizing capacity utilization of stations by estimating knock-on train delays”, Transportation Research Part B 41, 202–217.

Yung-Cheng, L., Yung-An, H. (2012), “Estimation of Single and Double-Track Capacity with Parametric Models”, Transportation Research Board 91st Annual Meeting