



ROYAL INSTITUTE
OF TECHNOLOGY

Simulation of rail traffic

—

Methods for timetable construction,
delay modeling and infrastructure evaluation

Hans Sipilä

Doctoral Thesis in Infrastructure

Stockholm 2015

TRITA-TSC-PHD 15-001 X
ISBN 978-91-87353-64-2

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

Abstract

This thesis covers applications and proposes methods for using simulation in a more effective way and also in a wider context. One of the proposed methods deals with delay modeling that can be used in a calibration process. Furthermore, a method is presented that facilitates the management of having timetables, infrastructure scenarios and delays as variables in simulation studies. The simulation software used in this thesis is RailSys, which uses a microscopic formulation to describe the infrastructure and train movements.

Timetable changes with respect to allowances and buffer times are applied on a real case (Western Main Line) in Sweden in order to see how the on-time performance is affected for high-speed passenger trains. The potential benefit is that increased allowances and buffer times will decrease the probability of train interactions and events where the scheduled train sequence is changed. The on-time performance improves when allowances are increased and when buffer times concerning high-speed trains are adjusted to at least five minutes in locations with potential conflicts. Increasing the allowance gives similar improvements in on-time performance as increasing buffer times. However, the former case means that the total run time increases, which can have a negative impact on the attractiveness of the trains in the passenger's point of view. Decreasing the allowance naturally caused a reduction in the on-time performance.

Although the aggregated on-time performance shows improvements for both directions, the individual on-time performance for trains highlights the variance inherent in the different train slots. Considerable improvements are observed for some of the trains. One drawback with this approach is that it can consume more space in a timetable at certain locations, hence other trains may need adjustments in order reach these buffer times. A check of a heavily aggregated on-time performance for other passenger trains as well as freight trains did not indicate decreased on-time performance.

Setting up simulations, especially in large networks, can take significant amount of time and effort. One of the reasons is that different types of delay distributions, representing primary events, are required in order to obtain conformity with reality if a real timetable and network is modeled. Considering train registration data in Sweden, the separation in primary and secondary delays is not straightforward. The policy of initiating cause reports is the indication of delay increase of a certain magnitude between two adjacent stations. Even though the reports would be consistent, they fail to capture all the smaller deviations from scheduled run times which are important.

The method presented uses the basic train registration data to compile distributions of run time deviations for different train groups in a network. The idea is then to reduce these distributions by different percentages and perform simulation runs and apply the root mean square error (RMSE) as a goodness-of-fit measure to assess how accurately the model predicts the observed outcome. Hence, the

purpose is to filter out the secondary delays from the registration data since these should be modeled in the simulation and not explicitly introduced. The results from the Southern Main Line case study show that a reasonable good fit was obtained, both for means and standard deviations of delays.

In Sweden, freight trains show a large variance with respect to established schedules, they can depart significantly ahead of as well as behind schedule. This behavior cannot be captured in a straightforward way. In conjunction with the simulation on the Southern Main Line, a method is applied which basically seeks to time-shift all freight trains ahead of their defined schedule so that the true initiation distributions can be used. In previous studies, early freight trains were modeled as if they were on time. The simulation results indicate that passenger trains are not significantly affected by this change in methodology. Naturally, the freight operations can be modeled more closer to real conditions. The findings are consistent with a field study performed in Sweden year 2009.

In addition to the already mentioned methods, which are applied on real networks, a method for reducing the uncertainties by making assumptions of future conditions is proposed. It is based on creating combinatorial departure times for train groups and locations and formulating the input as nominal timetables to RailSys. The dispatching algorithm implemented in the software can then be utilized to provide feasible, conflict-managed, timetables which can be evaluated, for example with respect to scheduled delays. Then there is a possibility to proceed with a study and perform operational simulations with stochastic delays on a subset of the provided timetables. These can then consequently be evaluated with respect to mean delays, on-time performance etc.

The large number of combinations, which can become large even in a relative simple case, can be sampled in order to get a manageable size for simulations. Considering the properties of synchronous simulation, complex situations on single-track lines can result in deadlocks. Hence, the number of realized timetables can be lower than what is setup prior to simulation. This should be considered when evaluating data and making conclusions.

To facility the use of the infrastructure as a variable in these type of studies, an infrastructure generator is developed which makes it relatively easy to design different station layouts and produce complete node-link structures and other necessary definitions. The number, location and type of stations as well as the linking of stations through single-track or multi-track sections can be done for multiple infrastructure scenarios. Although the infrastructure can be defined manually in RailSys, a considerably amount of time and effort may be needed. In order to examine the feasibility of this method, case studies are performed on fictive lines consisting mostly of single-track sections. This shows that the method is useful, especially when multiple scenarios are studied and the assumptions on timetables consist of departure intervals for train groups and their stop patterns.

Acknowledgements

First of all I would like to thank my family. Trafikverket (Swedish Transport Administration) and SJ AB provided funding for this research. I would like to thank my supervisors Bo-Lennart Nelldal and Oskar Fröidh 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. In addition, I would like to thank the other members of the reference group: Armin Ruge, Per Köhler, Åke Lundberg and Magdalena Grimm (Trafikverket); Marie Dagerholm, Dan Olofsson and Britt-Marie Olsson (SJ AB). Thanks to Olov Lindfeldt, Karl-Lennart Bång and Anders Lindahl for valuable comments regarding the thesis. Thanks also to Anders Lindfeldt, Jennifer Warg, Jiali Fu, Behzad Kordnejad, Hans E Boysen, Josef Andersson, Fredrik Hagelin and Markus Bohlin for all the stimulating discussions and their valuable inputs. The discussions on simulation methods with Pär Johansson, Magnus Backman and Johan Mattisson (Trafikverket) have been rewarding. Finally, I wish to thank fellow PhD students and other colleagues at the Department of Transport Science and Δ .

Stockholm, May 2015

Hans Sipilä

List of publications

Papers

- A Sipilä, H., 2010. Simulation of modified timetables for high speed trains Stockholm–Göteborg. Published in: Proceedings of the First International Conference on Road and Rail Infrastructure, Opatija, Croatia.
- B Sipilä, H., 2011. Calibration of simulation model on the Southern Main Line in Sweden. Published in: Proceedings of Railway Engineering, 11th International Conference, London, UK.
- C Lindfeldt A., Sipilä, H., 2014. Simulation of freight train operations with departures ahead of schedule. Published in: Computers in Railways XIV – Proceedings of the 14th International Conference on Railway Engineering Design and Optimization (CompRail 2014), Rome, Italy.
- D Fröidh, O., Sipilä, H., Warg, J., 2014. Capacity for express trains on mixed traffic lines. Published in: International Journal of Rail Transportation, Vol. 2, No. 1.
- E Sipilä, H., 2014. Evaluation of single track timetables using simulation. Published in: Proceedings of the IEEE/ASME 2014 Joint Rail Conference, Colorado Spings, USA.
- F Sipilä, H., 2015. A simulation based framework for evaluating effects of infrastructure improvements on scheduled and operational delays. Published in: Proceedings of the 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo), Tokyo, Japan.

Related publications not included in thesis

- Sipilä, H., 2008. Körtidsberäkningar för Gröna tåget: Analys av tågkonfigurationer (Run time calculations for the Green Train: Train configuration analysis). In Swedish. Technical Report, KTH.
- Nelldal, B-L., Lindfeldt, O., Sipilä, H., Wolfmaier, J., 2008. Förbättrad punktlighet på X2000: Analys med hjälp av simulering (Improving punctuality for X2000: Simulation analysis). In Swedish. Technical Report, KTH.
- Lindfeldt, O., Sipilä, H., 2009. Validation of a simulation model for mixed traffic on Swedish double-track railway line. Published in: Proceedings of Railway Engineering, 10th International Conference, London, UK.

- Sipilä, H., 2010. Tidtabellsläggning med hjälp av simulering: Effekter av olika tillägg och marginaler på X2000-tågen Stockholm–Göteborg (Timetable planning using simulation: Effects of supplements and allowances for the X2000 trains Stockholm–Gothenburg)). In Swedish. Technical Report, KTH.
- Sipilä, H., Warg, J., 2012. Kapacitetsanalys av Södra stambanan: Effekter av ökad trafik och ökad hastighet från 200 till 250 km/h (Capacity analysis of the Southern Main Line: Effects of increased traffic and increased speed from 200 to 250 km/h). In Swedish. Technical Report, KTH.
- Sipilä, H., 2013. Timetable generation on single track lines using combinatorics and simulation. Published in: Proceedings of Railway Engineering, 12th International Conference, London, UK.

Contents

1	Introduction	1
1.1	Objectives	2
1.2	Railway operating principles	2
1.3	Quality of service measures	7
1.4	Explanation of terms	8
2	Literature study	11
2.1	General railway operations	11
2.2	Capacity and Timetabling	13
2.3	Simulation methods	16
3	Rail traffic simulation methodology	23
3.1	Infrastructure	24
3.2	Trains and timetables	25
3.3	Simulation with delays	25
3.4	Discussion of the methodological problems	26
3.5	Limitations	28
4	Measures for improving on-time performance for high-speed passenger trains (Paper A)	31
4.1	Allowances and buffer times in timetables	31
4.2	Case study	32
4.3	Results	33
4.4	Conclusions	36
5	Method for calibrating primary run time delays (Paper B and C)	39
5.1	Background	39
5.2	Method	40
5.3	Results	42
5.4	Modeling freight trains ahead of schedule	45
5.5	Conclusions	47
6	Calculation of run times for the Green Train (Paper D)	49
6.1	Background	49
6.2	Calculation of maximal curve speed	50
6.3	Speed profiles	52
6.4	Train characteristics	55
6.5	Run time calculations	57
6.6	Conclusions	60

7	Multiple timetable and infrastructure analysis (Paper E and F)	61
7.1	Nominal timetables	63
7.2	Operational timetables	66
7.3	Stochastic simulations of operational timetables	70
7.4	Assigning delays	71
7.5	Infrastructure as a variable	72
7.6	Infrastructure modeling	74
7.7	Handling of deadlocks	76
7.8	Conclusions	77
8	Contribution of thesis	81
8.1	Methodological framework	81
8.2	Practical framework	82
8.3	Future work	82

1 Introduction

The transport performance in Sweden has increased with 35% for passenger transports (passenger-kilometers) and with 7% for freight transports (tonne-kilometers) over the last decade (Trafikanalys, 2015). Another study compares years 2008 and 2012 in Sweden and concludes that the number of passenger trains increased with 10–40% and the number of freight trains with 10% in total (Lindfeldt, 2009, 2014). Viewed in a global perspective, the number of passengers traveling with railway has increased with 23% over the two last decades (UNECE, 2014).

Sweden is experiencing increasing capacity constraints across the rail network, hence finding solutions that can meet this and expected future growth is important. The pressure increases on the existing rail network. Although it is a positive trend seen from an environmental perspective, several problems with rail services are highlighted with the increasing traffic load. More trains and increased capacity utilization lead to higher probabilities for delay propagation in the system. Operational characteristics common in Sweden, such as mixing trains with significant differences in average speeds on the same line, is an aggravating factor. A higher utilization of the infrastructure requires more maintenance in order to provide a sufficient level of reliability. In the long term new tracks and lines are required, but there is a need for actions and strategies that can improve performance on the existing network in the shorter term as well.

In the past, the whole rail network (train services and infrastructure) was operated and maintained by the national railway company. Today functions are split among several operators, an infrastructure agency, maintenance companies etc. Increasing demand for transportation affects commuter and regional train systems as well as long distance services. This can for example mean that the number of trains increases, the areas served by different train systems are extended and so on. Introducing new train stops on stations will improve the situation for some passengers. On the other hand this can also create problems to an already congested rail network, especially on the main lines. While the frequency of different train services is important, other transportation capacity measures should also be considered. This can involve longer trains and double-decker trains that can carry more passengers. Increased demand for freight transportation will transform to requests for more freight trains as well as longer and heavier trains.

One way of evaluating or predicting the performance of a railway timetable is to use simulation. It can correspond to both long, medium and short term planning. Assuming that the infrastructure, train runs and other parameters are properly modeled, a simulation can tell something about the expected delays and on-time performance. Alterations for one or several parameters followed by another simulation can then indicate whether an improvement can be expected or not compared with the initial case.

This thesis covers some of the applications that are important in simulations focused on timetable planning and proposes a methodology for expanded use of simulation in long and medium term planning. The research project is called *Timetable planning using simulation* and it is financed by the Swedish Transport Administration and Sweden's largest passenger train operator SJ AB.

1.1 Objectives

The main objective in this research project is to develop and improve methods for using simulation in the Swedish timetable planning process. Today this work is mainly done by the Swedish Transport Administration (Trafikverket) and, to a varying degree, in cooperation with train operators. Although softwares are used, the procedure is largely similar to manual planning. As mentioned earlier, increasing demand for more trains require reliable methods for predicting expected outcome of timetables.

The objectives include the use of example cases, both real and fictive, and show their application to timetable planning. This covers infrastructure and vehicle modeling, delay modeling and evaluation of simulation results. The presented methodology and findings should be applicable in the timetable planning process in Sweden, both for the long and medium term planning.

KTH has used the train traffic simulation software RailSys since year 2000 for capacity, infrastructure and timetable studies. The Swedish Transport Administration introduced RailSys in its organization in year 2006 and has gradually begun using it for capacity studies and also partly as a planning tool. Seen in this context, RailSys was therefore the primary choice of simulation tool also in this research project.

1.2 Railway operating principles

Many railways are operated on a timetable basis, meaning that train runs are scheduled in advance and should ideally be conflict-free. Typically a timetable is described by a time-distance diagram in which a train is represented by its stringline graph, a so-called graphical timetable. However, a train is in reality occupying a time channel around its time-distance line. A model that describes that time channel very exactly is the so-called blocking time model (see e.g. Pachl, 2002). Trains need a movement authority in order to operate, this is necessary for maintaining a safe separation between trains. Traditionally this authority is transmitted at discrete points, e.g. by lineside signals and/or transponders (balises) used as data points in an intermittent automatic train protection system (ATP). A train may only enter a track section, which in this case is the section between two successive signals, when it is not occupied by other vehicles.

Depending on the operated speeds and the track section distances, a train may need movement authority extending over several track sections in order to maintain safe separation at a certain speed. Thus, the blocking time of a track section is usually much longer than the time the train occupies the section. The length of a track section varies, on stations and municipal lines with dense traffic typically from a few hundred meters (sometimes even less) up to several kilometers in other parts of a network. Fig. 1 illustrates the principle with movement authority and track sections. In a graphical timetable, the time channel representation will naturally look different in the two cases. In an operational scenario the movement authority can of course be reduced, the speed must consequently also be reduced. If the movement authority is continuously updated, the fixed track section principle may not necessarily apply and therefore the safe operation distance between trains may also be reduced.

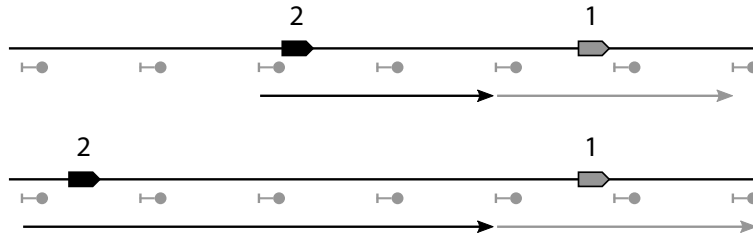


Figure 1: Principle for track sections and how the length of movement authority changes for higher speeds.

If we consider a railway line under normal conditions, the objective would be to schedule the trains to as far as possible avoid reductions in movement authorities of the type described in fig. 1. If all trains are running with approximately the same speed the traffic is homogeneous, the opposite is called heterogeneous. Scheduling trains with significant speed differences will consume more space in the graphical timetable, hence less trains can be scheduled. If we consider a longer section of railway line including stations, the average speed becomes important which takes into account the time it takes to brake and accelerate for scheduled stops as well as the standstill time.

Setting and releasing routes will also take some time, since the state of objects must be checked and changed if necessary. This is handled an interlocking system and the time for setting and releasing routes varies with different systems. Typically 10–30 seconds is needed for setting routes, less time for releasing. Representing the blocking times for all train movements in all locations on a railway line in a stringline graph results in a blocking-time diagram.

Operations on stations can induce further timetable limitations due to restrictions in the number of available tracks and restrictions in simultaneous train movements. Some are obvious, e.g. that two trains cannot simultaneously get movement authority on routes that cross each other at some point. Others have

to do with the fact that there may be requirements of a reserved safety distance (overlap) extending beyond the point of a movement authority. A typical example applies for train meets on single-track line stations with a configuration that do not allow simultaneous entry, thus affecting the timetable design.

Both during a timetable planning process and in real operations different train groups may have different priority values. A common practice is that slower trains give way for faster trains. This implies for example that slower trains must be moved aside to let faster trains pass (overtake) or meet. In real operations this may also be the case since a fast train is potentially sensitive to disturbances, i.e. a significant delay increase may occur if obstructed by other trains or in some other way. Additionally there may be operating policies stating that trains on time should not be obstructed by trains off time. If a good overall solution is pursued, it is not necessarily the best practice to hold back a late train.

Deviations from a planned schedule will almost always occur to a degree. Small deviations should be absorbed by allowances combined with buffer times. Allowance, also called slack or supplement, can be the difference between minimum run time and the scheduled one (run time allowance) or it can be an extra time added to a station stop (dwell allowance). Buffer times are added to the minimum headways (time spacing) between trains in a blocking-time diagram. Buffer times are of interest in all locations where a smaller delay on one or more trains implies a route conflict. The main purpose with allowances and buffer times is to avoid or reduce delays, or in other words to decrease the number and size of smaller delays (Rudolph, 2003). They can never compensate for all disruptions taking place in a railway system. How much allowance and how large buffer times are needed and in what way they should be distributed in a timetable is however not always straightforward.

Run time allowance may consist of several elements. A standard train category dependent allowance percentage can be included for all scheduled trains by default. Additional allowance can then be introduced and this can for example be distance based, i.e. longer running distance leads to more total allowance if measured in a time unit. It can also be used as a speed equalization measure on heavily utilized sections with a mix of train categories and stop patterns.

Although allowances and buffer times consume capacity, allowances also extends the total run time, they are often necessary in maintaining an acceptable on-time performance. One challenge in timetable construction is where and how to allocate this additional time and the amount of buffer that is needed at different locations in order to reduce the probability of relatively small disturbances leading to delay propagation. In his context a categorization is made into primary (exogenous) and secondary (knock-on) delays in which the first one refers to an exogenous event affecting one train and the second one to the knock-on effect if the delay propagates to another train, i.e. a delay caused by interaction between trains. The terms initial and reactionary delay are also used in some literature. Examples of events that cause primary delays are:

- Infrastructure and rolling stock malfunctioning
- Excessive alighting and boarding times of passengers and excessive freight handling times
- Reduced traction due to weather conditions

Considering a railway system with a conflict free timetable, meaning that there are no route conflicts caused by overlapping blocking (occupation) times, all knock-on events originate from at least one exogenous event. However, the chain of events leading to a specific knock-on delay for some train, the propagation, can be hard to trace exactly in a system with many trains and dense traffic. Under these conditions, one delayed train can cause delays to several other trains over a large area and a long period of time. The recovery time needed to return to scheduled operation is also an interesting parameter and relates to the reliability of a timetable.

How a timetable on a line or network should be designed is influenced by multiple factors. First of all there is a demand for the different train services and this will most likely vary during a day and over a year. The locations of single- and multi-track sections, stations and so on provide infrastructure conditions. Available vehicles and their performance define possible trip chaining schedules (vehicle circulation). Experience and analysis of historical data can give input for the allocation of allowances and buffer times. Transfers (connections) between trains are ideally designed with attractive passenger waiting times, i.e. not too long, but with sufficient margin in order to tolerate smaller or regularly occurring delays. Furthermore, the needs of freight transportations should also be considered. Consequently, a timetable construction process is often complex with all of these factors counted.

On a deregulated rail traffic market, different train operators apply for train slots according to their preferences. Since trains cannot use the same track sections simultaneously and there must be separation time between trains, the requested train slots can rarely be met for all operators and trains involved. A way to visualize it is to think that all requests are inserted in a graphical timetable, hence a train is represented by its nominal schedule. This can only be realized if no other train interferes with it. In order to obtain an operational timetable, the trains must be rearranged so that no occupational conflicts exist.

Fig. 2 shows three stringline diagrams with time on one axis and distance on the other. The first one represents requested train paths, hence it is a nominal timetable. In this single-track line example trains can only meet or pass on locations equipped with at least one siding (loop track). Removing conflicts by arranging the trains gives an operational timetable with bold stringlines, the nominal train paths (thin stringlines) are left for comparison. In real operations delays will occur, meaning that there will be deviations in the realized operational timetable. The realized operation is shown with bold stringlines and the planned operation with thin stringlines.

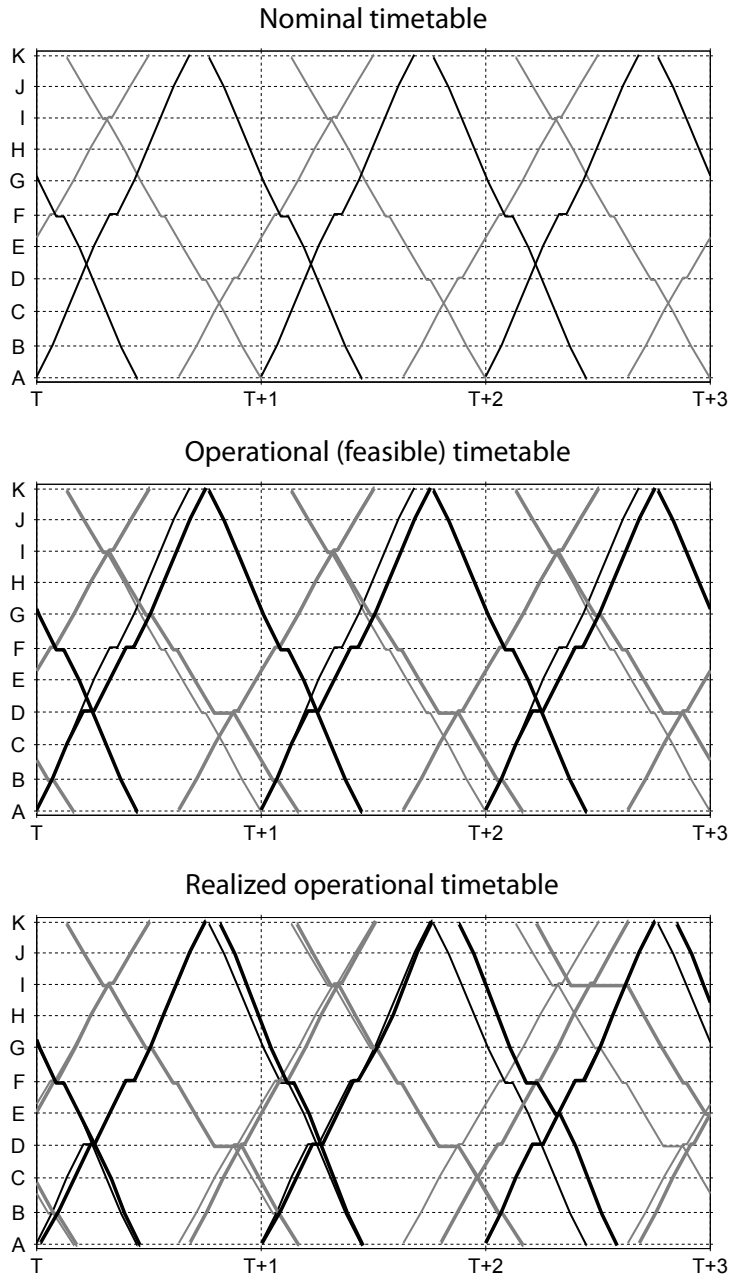


Figure 2: Example of a process starting from requested times filed in by train operators, which can be said to constitute a nominal timetable. This is followed by conflict-management to obtain an operational timetable. Multiple realizations (day by day) of the operational timetable provide information on the reliability.

Operating a railway according to the second diagram in fig. 2 is equivalent to a scheduled or structured operation. In theory, all trains can run on schedule without interference from each other as long as no primary delay occurs. Running trains according to the first diagram, i.e. without prior conflict-handling, is referred to as an unscheduled or improvised operation. It gives flexibility since freight trains can be operated on a day-to-day demand and accumulate freight until a decision is made to dispatch the train. However, as the number of trains on a line or network increases this type of operation can be difficult to manage. Scheduled operation is the normal case in Europe. Unscheduled operation is common in North America with a high percentage of freight trains. A scheduled operation that suffers significant delays partly be viewed as an unscheduled operation until the scheduled operation is restored.

1.3 Quality of service measures

One common quality of service measure is on-time performance, also referred to as punctuality. It is measured as the percentage of the trains that arrive or depart at a location with a delay less than a certain threshold value. The threshold values vary between different countries, train types and operators. On-time performance values are sometimes heavily aggregated for multiple train groups and can for example refer to terminal station arrival values for all long-distance trains in a country. Using one on-time performance threshold value provides one point on a cumulative distribution. Using a sufficient number of different threshold values gives the cumulative distribution, hence providing more insight.

The mean delay refers to taking an average of the deviations from a scheduled time. Mean values are sensitive to extremes (outliers) relative to the majority of deviation values, e.g. a mean delay value is skewed upwards by a small number of trains with a high delay although the majority of trains would have a delay lower than the mean delay. A more useful value is obtained if a cut-off is done at some delay level meaning that extremes are not considered. Alternative measures can for example be to take the mean delay on all late trains, not including trains on time. A study reported in Warg and Bohlin (2015) highlights the importance of assessing the full arrival or departure distribution instead and not only represent all registrations with one average.

Including the standard deviation as a measure captures the dispersion, which is important since it indicates whether deviations are systematic or spread at some measurement point in a network. Hence, it is an expression for the precision in the operational traffic. The median can also be useful and complement other measures. The influence of outliers can be reduced, e.g. by using the median absolute deviation.

Studying these measures en route can highlight sections where delays increase or decrease. This information is useful in timetable planning and gives some guidance for implementing adjustments in order to improve conditions. Peterson

(2012) gives an example of analyzing the individual train performance for two high-speed train en route. If a delay grows gradually over the whole train run it may be an indication of that regular smaller delays occur along the route. Increased allowance may counteract in this situation. There may also be locations where a significant delay increase takes place, hence indicating that a regularly occurring event is frequent. This can for example be a change in the sequence trains are processed. Increasing buffer times at this location may improve the situation.

Other approaches giving further insight involve tracking of event chains and in that way getting a more comprehensive picture of the consequences caused by specific exogenous events. Margins allowing delay recovery influence the propagation of knock-on delays, thus making this type of evaluation difficult on data of low resolution regarding the number of measurement locations. In a simulation environment this type of approach is done more easily since the input is clearly defined and sufficient output data is available.

Timetable robustness is a term that refers to the ability of recovering from smaller delays and prevent delays from spreading. It can be applied on timetables prior to execution and on performance after a period of execution. Depending on the specific robustness measure used, it takes into account parameters such as allowances, headways (buffer times) and heterogeneity. Andersson et al. (2013b) propose a measure that can be used to evaluate a timetable with respect to points where trains are planned to enter a line or overtake another train. These so called critical points can be used to identify weaknesses in a timetable. The measure consists of three parts: the available runtime margin for the operating or overtaking train before the point and for the entering or overtaken train after the point, as well as the headway margin between the trains in the point. The measure is named robustness in critical points (RCP).

The frequency of passings affects capacity and is closely linked to delay sensitivity. This holds for single-track meets as well, but these must take place regardless of homogeneous or heterogeneous traffic. Therefore this parameter can be used as a predictive indicator of potential outcome. However, as mentioned earlier, passings and meetings imply scheduled delays for some trains. These can however also be used for reducing delays in real operations. Several of the measures mentioned may also be reported per distance to facilitate comparisons, e.g. delay increase per 100 km.

1.4 Explanation of terms

Table 1 explains most of the terms used in this thesis.

Table 1: Explanation of terms used in the study.

Station	Location where trains can switch between tracks.
Line (section)	Tracks between two adjacent stations
Train dispatching	Authorizing and implementing the sequence order of train movements, typically based on priority rules. In simulation models this is handled by an algorithm.
Allowance/supplement	Additional time introduced as added percentage of minimum run time or as a time value.
Buffer time /margin	Additional temporal distance between two trains, added to minimum headway.
Minimum headway	Minimum temporal distance between two adjacent trains.
Delay/disturbance	An event that disrupts planned operations. A delay is also an output of such an event.
On-time performance	Percentage of trains that e.g. have less than or equal to X minute delay, also called punctuality
Primary delay	Result from an initial/exogenous event, e.g. caused by infrastructure or vehicle malfunction
Secondary delay	Propagating (knock-on) through train interactions, originates from a primary delay.
Entry/initial delay	Used in simulation to delay train initiations.
Dwell delay	Used in simulation to model variance in stop times.
Line delay	Used in simulation to model variance in run times between stations, also called run time extension.
Minimum run time	Technical minimum run time between two locations including scheduled stop times.
Nominal run time	Minimum run time, can include allowance.
Nominal timetable	Timetable where trains have nominal run times and paths, not conflict-managed.
Scheduled run time	Run time in conflict managed timetable.
Scheduled delay	Difference between nominal and scheduled run time, also called scheduled waiting time.
Operational timetable	Conflict-managed, feasible, timetable.
Static performance	Performance measures for a timetable
Dynamic performance	Performance measures for a timetable operated in reality or in a simulation, exposed to delays.
Cycle	Simulation time period repeated multiple times in stochastic simulations.
Deadlock	Situation in simulation where trains block each other, hence the operation cannot proceed.

2 Literature study

2.1 General railway operations

Mattsson (2007) reviews some possibilities of analyzing train delays and their relationships with capacity utilization. Analytic, statistical and micro-simulation methods for delay analysis are presented and some of their advantages and drawbacks are described. The analytic methods are often using elements from queuing theory and optimization. They can lead to mathematical problems that may be time-consuming to solve. Compared to simulation models, the amount of input data may be small and an exact timetable is not always necessary. This makes them useful at a strategic planning and design stage.

Micro-simulation offers a useful approach in modeling interactions between trains, i.e. the propagation of knock-on delays initiated by exogenous events (primary delays). Defining input data, e.g. a track layout and corresponding timetable, can take a significant amount of time. Although a high level of preciseness can be achieved, the uncertainties introduced by making assumptions of future conditions should be addressed. Furthermore, there is a risk of losing resolution in general tendencies if only one specific setup is studied.

Statistical regression offers a possibility to estimate equations by which the amount of primary delays can be predicted based on factors such as track and rolling stock maintenance status, traffic volumes and weather conditions. Statistical analyzes of empirical data seem to be the only realistic way of modeling the occurrence of primary delays. Regression can give an opportunity for establishing empirical relationships between capacity utilization and knock-on delays for a given level of primary delays (see Gibson et al., 2002).

Olsson and Haugland (2004) discuss influencing factors on train punctuality and some applications of this. Examples of some analyzed punctuality factors are infrastructure capacity utilization, cancellations, operational priority rules and temporary speed restrictions. Data regarding both delays and the influencing factors have been used with the aim to identify how the studied factors influenced the punctuality. The study focused on commuter trains in the Oslo area and long distance trains on one line. Number of travelers, capacity utilization and temporary speed restrictions are found to have a negative correlation to punctuality. Cancellations have a positive correlation. Several of the factors seem to have threshold values. Delays under these threshold values are absorbed during normal conditions, high delays remain in most cases until the final destination.

Differentiating and quantifying primary and secondary delays at a high time resolution level is often difficult and depends how and on where train registration data is collected. If the resolution is low, the differentiation and causal tracking of delays becomes difficult and dependent on assumptions. An interesting approach for differentiating and quantifying primary and secondary delays is presented in

Labermeier (2013). In this Swiss study, train registration data for station arrivals and departures together with signal passage and aspect changes is available on a time scale of seconds. Their results show that secondary delays on stations contribute most on days with a low punctuality. An explanation for this is that connections play an important role in Swiss passenger train timetables. The findings can be used for constructing more robust timetables that would decrease the propagation of delays.

Eliasson and Börjesson (2014) argue that timetable assumptions are of crucial importance in railway investment appraisals. Explicit principles are needed for timetable assumptions in order to compare appraisals of railway investment with each other, otherwise cost-benefit analysis can be influenced, e.g. stakeholders' strategic behavior. The current practice is likely to exaggerate appraisal benefits. An increase in capacity can be used to increase frequency, shorten travel times, reduce delays by improving the operational conditions or a combination of these. Consequently, assumptions about future timetables become important when appraising capacity investments. The assumptions are equally important in the do-nothing scenario. The timetables must be derived considering demand and supply as well as current institutional settings and capacity constraints.

Numerical examples are used to illustrate how the outcome of a cost-benefit analysis becomes almost arbitrary when an explicit principle is missing for defining timetables in the scenarios. The number of trains per hour varies between maximizing consumer surplus, producer surplus or total social benefits. If two or more operators are involved their objectives may be quite different and they have incentives to state a future timetable different from their real intentions. A public operator can work to maximize consumer surplus whereas a commercial operator tries to maximize profits. Considering that railway investments often represent huge public spending commitments it is crucial that appraisal methods are reliable including an explicit and objective principle for timetable assumptions.

Milinković et al. (2013) suggests an approach based on an artificial neural network model for calculating arrival delays at stations in a network. A case study is performed using historical delay data as target data and predefined input data for each train. This is applied on one station in the Belgrade area in Serbia. On-time performance is low due to sections with temporary speed restrictions induced by poor track maintenance. Even though the utilization of railway line capacity is relatively low, the timetable is not able to compensate for a large percentage of delays. For training of the neural network model and the linear regression model for delays four factors are chosen. These are train category, arrival time, infrastructure influence and traveled distance.

Results of the neural network model and the linear regression model are compared to corresponding sample of real delays. This shows a significant correlation between actual delays and the results from the two models, of which the neural network model show better correlation. Goodness of fit tests are used to check whether the set of values are from the same population. This is fulfilled for the

neural network model but not for the regression model. Conclusions made is that the neural network approach can give good results in predicting delays in systems with large disturbances. It must however be tested on other stations in the same or other networks. It may ultimately be used as a complement in simulation models.

Marinov et al. (2013) concentrates on organizing, planning and managing train movements in a network. Three management levels – strategic, tactical and operational – are presented. The main difference is the time horizon within which they are applied. Strategic level encompasses long term planning and deals e.g. with new infrastructure and acquisition of new resources and technologies. Timetable construction is typically done on the tactical level which deals with medium term planning. The operational level is for short term planning and deals with implementing timetables and service commitments on a day-to-day basis to achieve the established goals.

Some basics of decision support systems and their application to train rescheduling are covered. Despite some technological developments, the rescheduling process is to a large extent handled manually by train dispatchers. However, rescheduling with conflict resolution is a difficult problem to solve and becomes computationally intensive with increasing number of conflicts and longer prediction time horizons. This process needs also frequent updating due to the dynamic environment in which it takes place.

2.2 Capacity and Timetabling

An application of a heuristic timetabling algorithm on a mesoscopic infrastructure description to the Train Timetabling Problem for large-scale networks is presented in de Fabris et al. (2014). Some of the drawbacks of macroscopic modeling are overcome while some of the benefits of microscopic modeling become available by using a mesoscopic description of infrastructure. In general, a significantly higher accuracy is possible compared to a macroscopic definition. Minimization of overall penalties paid for trains not inserted in the final timetable or for differences between realized and desired schedules for each train is one objective of the heuristic approach.

A second objective is that the timetables should be acceptable by human planners and applicable to real-world situations. A case study is presented for the North-East part of Italy including over 120 stations and both single and double track lines. This is performed under different demand conditions. A conclusion made from this is that the software is able to compute an acceptable solution for the network in a few minutes. A set of these timetables can be used as input to operational microsimulation.

A decision support framework for strategic railway capacity planning is presented in Lai and Barkan (2011). One aim is to help capacity planners determine how to allocate funds optimally. It consists of three modules; an alternative generator

which lines up possible expansion options with corresponding costs and capacity effects, an investment selection model which points out what parts of the network need to be upgraded and an impact analysis module evaluating the trade-off between capital investment and delay cost. The authors underline that operational options should be considered first because they are generally less expensive and can be implemented faster than building new infrastructure. This type of approach can be more applicable for network capacity planning than micro simulation, especially considering the strategic level planning horizon. Parts of the presented framework are based on a parametric railway capacity evaluation tool discussed in Lai and Barkan (2009).

Three types of capacity expansion alternatives are handled in the alternative generator: adding passing sidings, intermediate signals (shortening block lengths), a second main track. The investment selection model is formulated as a mixed-integer network design model, it determines the best set of investment options for capacity expansion provided that the level of service remains the same as the current conditions. The impact analysis module gives a net cost, delay cost and benefit for links subject to capacity expansion. The outcome may point out that some of the investments are not cost effective. Two case studies are analyzed to demonstrate the use of the decision support framework. The authors propose an expansion of the investment selection model making it possible to account for stochastic future demand or multiperiod decision making.

Fransoo and Bertrand (2000) presents an aggregate model that can be used to single out investment alternatives in the infrastructure, specifically passing constructions (sidings). The model aims to provide the user with information about the ranking of various alternatives and relate it to the theoretical capacity. The main reason for using a high level of aggregation is to overcome the tedious and time-consuming process if a detailed model is used. Referring to that nearly the entire passenger network in the Netherlands is double tracked, waiting times due to train meets are not a concern for the timetable planners. Since the tracks on a double track line can usually be divided in two directions it is sufficient to model only one direction of traffic.

The theoretical capacity is determined under some assumptions. A maximum of two kinds of trains on a section are handled, the trains are scheduled in alternating order to get the worst case with respect to timetable heterogeneity. This means that a local train is followed by an express train which in turn is followed by a local train and so on. An example line is used to compare results from the aggregate model with a more detailed analysis. The ranking order of alternatives show agreement comparing theoretical capacity whereas the increased riding time measure did not fully coincide.

Lindfeldt (2011b) investigates double-track railway line capacity and presents a combinatorial model – Timetable Variant Evaluation Model (TVEM) – for generating multiple timetables for different mix of heterogeneous traffic. In this model both infrastructure and timetable are modeled as variables. Train patterns are

scheduled asynchronously in order of their priority, hence the scheduled delay (scheduled waiting time) tends to increase for train patterns the later they are placed in the timetable. To capture the effect of varying distances between adjacent overtaking stations the infrastructure is modeled stochastically using Weibull distributed distances with different shape parameters. Three different operational cases with mixed traffic are analyzed. Conclusions made are that the influence on capacity from the infrastructure increases with speed difference and frequency for passenger trains, whereas the importance of the infrastructure design decreases when traffic is more heterogeneous.

The methodology and idea of TVEM is expanded for single-track line analysis in Lindfeldt (2011a), using the same asynchronous principle for timetable generation. Train patterns are created in a strict priority order and if already inserted they cannot be moved. This has the advantage that, depending on the setup, several timetables can be generated for long lines which share characteristics such as number of trains per hour. However, there is a relatively high probability of getting unevenly distributed scheduled delays between train patterns due to the asynchronous principle.

Forsgren et al. (2011) present a prototype tool – *the Maraca* – for non-periodic timetabling. The tool takes an already existing schedule and optimizes it with respect to resource conflicts. The purpose is to aid the timetable constructor by minimizing the resource conflicts in his or her draft timetable. Generating timetables with the tool is described as an iterative and interactive process. Many so called soft constraints are not included in the model and the solutions given by the tool need to be judged by a timetable constructor. While timetables in some countries can be described by modeling an hour of typical traffic, the shortest period of time that makes sense for a Swedish timetable is a whole day. Because of this, the use of the Periodic Event Scheduling Problem is not considered suitable, instead the presented model is more similar to general scheduling models.

A case study is performed to check the validity and the scalability of the model for its intended use. The intention was to regularly receive data from a timetable constructor, perform calculations with the tool and feed results back to the timetable planner for inspection. In the first set of test runs, the number of conflicts is minimized. It shows that letting the tool work on a problem for up to 15 minutes always resulted in a timetable that was most significantly improved compared with the draft regarding resource conflicts. In the second set, the number of conflict seconds is minimized instead. In these test runs it was more clear that a combination of things affects the capability of the model to reduce resource conflicts, e.g. the number of conflict seconds in the draft timetable and how close to each other trains in the draft are. A conclusion made based on the case studies is that the Swedish Transport Administration would benefit from introducing mathematical optimization in the timetabling process.

2.3 Simulation methods

Siefer (2008) gives an overview of railway simulation techniques and highlights the benefits of using operational simulations, e.g. for assessing infrastructure expansion alternatives and verifying timetables beforehand. Macroscopic models describe a network as a directed graph of simple nodes and links, representing stations and lines with attributes. Microscopic models describe a network in high detail and allows evaluation of train interactions on a detailed level. Synchronous models can simulate all events in a network in short time steps, whereas an asynchronous models initiates trains successively according to priority. As in many other computer models, the accuracy of the simulation results are strongly related to the exactness of the input data and the consistency of the algorithms used. Synchronous simulations are, especially with heavy traffic, prone to deadlocks on single track lines and to a certain extent in complex stations.

Differences between macroscopic, mesoscopic and microscopic infrastructure modeling and the pros and cons are discussed in Gille et al. (2008). Macroscopic models are well suited for long-term planning when average values for train running times are sufficient but are unable to describe transient train dynamics. They represent the infrastructure with a high level of abstraction where nodes and links contain aggregated information. The modeling of station operations and associated internal dependencies, e.g. route exclusions, cannot be fully described on a macroscopic infrastructure.

A microscopic model allows for a more realistic description of operations where nodes represent railway equipment objects and links has attributes such as length, speeds, gradients etc. However, computational inefficiency can prevent their application to large-sized networks, especially if the time factor is important. A mesoscopic infrastructure description offers a compromise in between the other two modes, e.g. to limit the complexity of the model or if detailed infrastructure data is unavailable. This allows for the definition of station routes and their connection to line sections and route exclusions regulating movements that cannot take place simultaneously.

Pachl (2007, 2011) discusses why deadlocks occur in synchronous railway simulations. A deadlock is a situation in which a number of trains cannot continue along their path at all because every train is blocked by another one. In asynchronous simulation, train groups are scheduled (inserted) in order of their priority and in order to get a timetable without scheduling conflicts some train paths have to be postponed. The situation is completely different in synchronous simulation in which the processes of railway operation are simulated in real time sequences. By introducing delays deadlocks may be produced since there is no parallel timetable processing during a running simulation. Even though a control logic can avoid simple forms of deadlocks, complex infrastructures and bidirectional operation are prone to deadlocks.

Two approaches for avoiding or reducing the occurrence of deadlocks are discussed. The basic idea of Movement Consequence Analysis (MCA) is to analyze which consequences will necessarily result from train movement for the further sequence of train movements. Primary consequences are train movements that must be made to enable a train to enter a section. Train movements that must be made after a train has entered a section are called secondary consequences. By creating logical tree structures deadlocks can be detected since they lead to closed loops. MCA can handle complex situations but it leaks flexibility since the usage of station tracks has to be fixed in advance over the entire train path. Dynamic Route Reservation (DRR) follows the principle that a certain number of track sections ahead must be reserved before a train may be authorized to enter a track section. The reservations are stacked like routes in an interlocking and describe the train sequence for a deadlock-free operation. In rare situations a deadlock may still occur.

White (2005) discusses different ways of simulation analysis and the effect it has on forming conclusions based on the simulation output data. The effect of infrastructure on traffic may for example not be easily associated with traffic conditions. This could partly be overcome if care is taken in preparing input data and an appropriate output analysis methodology is used. A commonly used simulation process is heuristic. Following the analysis of the simulation output, the infrastructure is modified or expanded in a way to meet a certain performance level under operational conditions. After a new simulation and subsequent analysis the process can be further repeated until an acceptable result is achieved. However, the author points out that this process can lead to incorrect solutions.

The two general philosophies of railroad operation are outlined, the improvised and structured operation. The author points out that understanding the infrastructure planning process used with structured operation can also be useful in developing infrastructure requirements for improvised operation. If no additional disturbances are introduced, the delay output of a simulation reflecting improvised operation should represent the inability of the infrastructure to accommodate the traffic. However if the traffic arrangement is altered, e.g. by changing departure times for one or more trains, the delay output can also change.

Simulation output data generally include several values from which other relationships can be prepared. Comparing different numbers do not necessarily provide conclusive results, some data can simultaneously provide conflicting results. Care must be taken in interpreting results to avoid making wrong conclusions. Root cause analysis can give insight about infrastructure adequacy but is not straightforward and often time consuming.

A simulation study investigating the impact on capacity and robustness from different infrastructure layouts and timetable variants with heterogeneous traffic is outlined in Lindahl (2002). Some of the conclusions is that introducing overtaking stations (passing loops) is an effective way to facilitate mixed traffic with both faster and slower trains, thereby increasing capacity on double-track lines. To get

the most benefit, they should be located centrally between adjacent stations. Equipping stations with switches that allow for higher speeds in diverging mode in conjunction with longer sidings (loop tracks) enhances capacity. This makes it also more flexible to operate long freight trains, thus increasing the transport capacity. A significant delay increase is observed when the temporal space is utilized to two thirds, while reducing the time between trains and observing decreased on-time performance and robustness.

Measures for improving on-time performance for high-speed passenger trains on the Western Main Line between Stockholm and Gothenburg are discussed in Nell-dal et al. (2008). Several scenarios are simulated in RailSys to give an idea of the impact on on-time performance if the level of primary delays is decreased for high-speed trains. Historical data is used for compiling delay distributions and these are further categorized into operator, infrastructure and vehicle related events. Results showed that even if the delays in all categories were reduced by 25–50%, on-time performance barely reached 90% at the end stations (trains with a maximum delay of five minutes). The conclusions were, i.a., that improvements are necessary for other train categories as well to reach up to a 95% on-time performance for the high-speed trains.

An experimental design setup was used in Lindfeldt and Sipilä (2009) to both calibrate and validate a simulation model on the Western Main Line. In this study registration data and timetable represented the same period. Mean and standard deviation of delays for different train groups are used as calibration and validation measures. Delay distributions representing primary run time extensions, dispatching priority settings and percentages of allowance that can be employed to reduce delays are used as variables in the setup. A good fit is achieved compared to operational outcome. Calibration and validation data sets were divided according to levels of exit delays, thus representing days with low and days with high delays.

Landex (2010) presents methods for estimating scheduled waiting time in railway networks by simulation. Waiting times for both trains and passengers are examined meaning that whole journeys with transfers between trains are considered. Hence, studying a specific journey with transfers can give that waiting times are low for the trains involved and simultaneously high for passengers due to long transfer times. The article mentions two models for calculating scheduled waiting time, the Danish SCAN model (Strategic Capacity Analysis of Network) and the North American TPC model (Train Performance Calculator). The key difference between these models is that the first one examines randomly generated regular interval timetables whereas the second one examines timetables where all trains are operated randomly.

It is pointed out that improving a timetable without considering its operational performance, e.g. by constructing a timetable with short transfer times thereby decreasing the scheduled waiting time for passengers, can give an over-optimized timetable in which small delays result in broken connections for passengers. Dif-

ferent timetables can be simulated with normal operational delays in order to estimate additional scheduled waiting times for passengers.

A field study of the Red Line of the Massachusetts Bay Transportation Authority is presented together with a simulation model (SimMETRO) in Koutsopoulos and Wang (2007). Different control strategies to improve the operating efficiency are tested and compared. Evaluation methodology and measures of performance parameters are discussed and the importance of making both a calibration and a validation process is emphasized. Methodology for modeling a calibration process as a multi-variate optimization problem and solving it with SPSA algorithm is discussed in Koutsopoulos and Wang (2011). This methodology is applied for the previously introduced simulation model (Koutsopoulos and Wang, 2007) and results show that the calibration process improves the parameters and refines the input. The algorithm used is shown to be efficient.

A simulation study, in which a hypothetical single-track line is modeled, is presented in Dingler et al. (2009). The software used for simulations is Rail Traffic Controller (RTC). Combinations of three different types of freight trains and one passenger train with different percentages of each train type are simulated. One objective is to investigate what aspects of heterogeneity have the most pronounced impact on delay. The principal metric used for comparison is average delay. In cases with only freight trains, heterogeneity in top speed, power to weight ratio and dispatching priority is varied separately for percentages of two selected freight train types. Results from simulations show that dispatching priority has a greater impact on delays than speed or power. Adding passenger trains gives higher delays compared to adding the same number of freight trains. This is due to both higher priorities and speeds.

Sogin et al. (2011) investigate capacity on single-track lines with the addition of both freight and passenger trains. Simulation software RTC is used to evaluate a representative fictive route with varying train density. Simplifications are made to improve comparison of the effects of key variables regarding traffic composition and passenger train speed. Some randomization of freight train departures are introduced. A homogeneous condition with a composition of 100% freight trains is defined as a base case, although the number of trains vary. Comparisons of simulations, where the total number of trains is constant, show higher delays in the heterogeneous cases. The more passenger trains operated, the higher the variation in freight train delays. Higher passenger speeds will increase delays although the marginal increase in speed has a diminishing factor on the delays of freight trains when the network becomes saturated.

The relation between capacity and percentage of double track, on an otherwise single track line, is studied in Sogin et al. (2013). An incremental transition from single to double track is achieved by adding a second main track between passing sidings (stations). Eight different traffic levels are studied starting from 8 trains per day up to 64 trains per day. These levels are simulated on 14 different track configurations, from a line with only passing sidings and up to full double track.

Simulations are performed with Rail Traffic Controller (RTC). A response surface model is developed with the aim to be able to predict the capacity of a line as a function of the amount of double track and the minimum level of service. Evidence from this study suggests that delays will decrease linearly for each marginal section of double track installed and it occurs for each of the applied traffic levels (freight trains). The linear reduction in delay is greater for the higher traffic levels than for the lower ones.

The strategy for installing double track is to pick a number of points on the line and build-out in both directions from these points. Another strategy studied is to build-out from the end points on the line and inwards. One objective is to investigate if the relationships between slope and intercept parameters can be formulated in an equation that could predict delay for a given double track percentage and traffic volume. Regression analysis is used to estimate exponential delay-volume relationships. Relating the percentage of double track to volume shows that the incremental capacity gained from each section of double track added increases as more double track is added to the line.

A comparison of how delays vary with respect to headway in both an event-driven mesoscopic and a time-driven microscopic simulation model of the same infrastructure is presented in Quaglietta et al. (2011). It is applied to a Mass Rapid Transit case study. The first train experiences an initial delay at a station and this affects, depending on the headway, a varying number of the subsequent trains. Fourteen different scheduled headway values are used from 90 to 480 seconds. Results show greater differences between the mesoscopic and microscopic models for shorter headway values whereas longer values give smaller differences. The mesoscopic model underestimates the arrival delays by 40% in the case with minimum headway. When a headway of 360 seconds is used the delays are underestimated by 5%.

Shorter scheduled headways mean that the probability of train interaction increases in case of disturbances. Modeling transient train dynamics becomes important when trains interact with each other. The mesoscopic model cannot capture this effect in a good way and this is the main reason for why the difference between the models increases with shorter headways. A point made in this paper is that a hybrid methodology could be used, i.e. some areas in an infrastructure model can be modeled with microscopic detail level whereas other areas can use a mesoscopic level. What type of model description is used can also be determined by delay types, i.e. if only smaller initial delays occur in a simulation cycle the mesoscopic model is used, otherwise a switch can be made to the corresponding microscopic model for the entire simulation cycle.

Marinov and Viegas (2011) describes a mesoscopic modeling approach that can be used for evaluating freight operations from a tactical viewpoint. A rail network can be separated into its components such as lines, yards, stations etc. The components are considered as interconnected queuing systems that interact and influence one another, hence capturing the impact of freight train operations in

a network. A case study is made that covers part of the rail freight network in Portugal focusing on the cost difference between improvised (unscheduled) and structured (scheduled) operation of freight trains.

Some of the mentioned shortcomings of the improvised operation are that it is not explicitly focused on customer needs, freight yards require a greater storage capacity to handle traffic variability and uncertainty, a greater number of road crews and road locomotives are required to satisfy customer demands. The structured operation, on the other hand, requires integrated, reliable, detailed and efficient operating plans. If the operations have difficulties to fulfill the schedules, customer demands are not met and the service is seen as unreliable.

The rail network is described by a set of Work Centers and Storage Areas using the Simul8 computer package for event-based simulation. Results from the case study indicate that the more structured and scheduled the network freight train operation is, the lower the yard queue becomes. Also, the more disorganized the freight train movements becomes, the larger the yard queue grows. The number of served freight trains decrease with greater deviations in schedules. From a monetary viewpoint, the effort should be towards a controlled fixed freight train service instead of adding extra yard personnel to handle significant variability.

Confessore et al. (2009) describe an approach for estimating the commercial capacity of railways. They combine a compression method defined in the UIC Code 406 and implemented by an algorithm with a discrete event simulation model for testing the robustness of the solutions. The approach is tested on data from a line in Italy characterized by substantial shares of both freight and passenger trains. Results are calculated for several minimum time spacing (headway) values and comparisons made for the capacity on identified line sections and the whole line. The difference falls in the interval 30–50%, the commercial capacity for the whole line being lower than for individual sections. The variability is around 50% between using six or one minute as time minimum spacing between trains.

Hwang and Liu (2010) describe a simulation model for delay evaluation, validated on a regional railway in Taiwan. The main purpose is to evaluate the impacts of primary (first) and knock-on delays on the modeled timetable. Dispatching measures can reduce dwell and running times for delayed trains. Validation is done by comparing actual and estimated delays at selected stations. The case study shows that the simulation model can be used for evaluating knock-on delays reasonably well. However, the studied timetable is relatively homogeneous and there are no high-speed or freight trains.

Kunimatsu et al. (2013) presents a method for estimating the effectiveness of turn tracks from a passenger viewpoint. Turn tracks (switches/turnouts) are needed to limit the extent of a disruption since they allow operations on both sides of a blocked section. However, installing and maintaining switches is associated with significant costs meaning that deciding locations for additional switches should be based on a thorough investigation.

A comparative case study is performed where passenger flows are compiled from origin-destination (OD) data. Disruption data with frequencies, locations and durations are used as input for simulations. Rescheduling patterns are specified for the different scenarios. Passengers can use detour connections, with varying number of transfers, to reach their destinations. A disutility function is adopted consisting of needed time to arrive at destination, experienced waiting time for trains and congestion etc. The results show that this method could be useful for investigating this type of infrastructure improvements.

3 Rail traffic simulation methodology

Simulation is the imitation of an operation of a real-world process or system over time, which should be as close as possible to its real-world equivalent (Abril et al., 2007). Simulation offers a way of evaluating different scenarios and actions prior to decisions are made concerning a real system. In rail traffic applications, simulation can be used to predict the outcome of a timetable when exposed to different types of delays.

If the outcome of a simulation shows that requirements are not met, changes can be made in the timetable and a new simulation performed to see if the measures undertaken are effective. In a wider perspective, infrastructure expansions and improved vehicle performance can for example be considered. Consequently, the conditions for timetabling will change as well. These measures can also lead to significant improvements regarding the incidence of exogenous delays having a positive effect on the expected outcome. Fig. 3 outlines the main elements in rail traffic simulation.

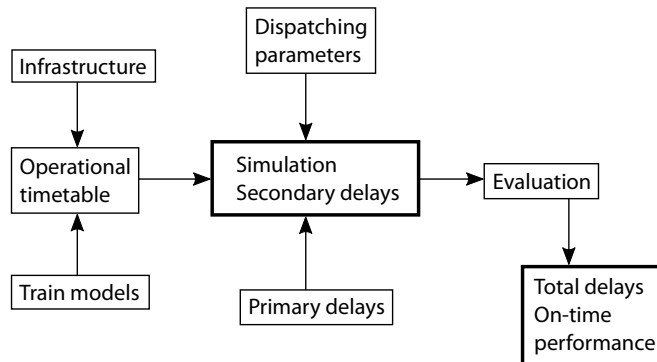


Figure 3: Rail traffic simulation methodology.

The simulations described in this thesis are done with the microscopic simulation software RailSys. It was developed by the Railway Group (IVE) at the University of Hannover and later at the University of Braunschweig. The commercial part is handled by Rail Management Consultants (RMCon). A description of RailSys is presented in Bendfeldt et al. (2000). Another microscopic simulation software is for example OpenTrack, developed at ETH Zürich and marketed by OpenTrack Railway Technology in Switzerland (see Nash and Huerlimann, 2004). An example of a simulation study, where different options for increasing the capacity in and in the proximity of a large terminal station, utilizing OpenTrack is reported in Kantamaa et al. (2013).

A software widely used in North America is Rail Traffic Controller (RTC) developed and marketed by Berkeley Simulation Software. Also this software uses a microscopic description and is used extensively by the U.S. rail industry. Both RailSys and OpenTrack are in principle timetable-based reflecting European conditions. RTC is primarily non-timetable-based following the unscheduled operation characteristics common in North America. Pouryousef and Lautala (2014) present a case study on shared-use U.S. rail corridor and discuss some similarities and differences of RTC and RailSys regarding infrastructure, train and timetable modeling.

All of these three softwares use synchronous simulation. This means that the simulation is driven following the real time sequences by flexible time intervals in an event-driven simulation or by fixed time intervals in a time-driven simulation. An event is a time-related occurrence where the status of a train changes. Some of the limitations are discussed in section 3.5.

3.1 Infrastructure

The infrastructure can be modeled on a macroscopic or microscopic level depending on the effects a specific study aims to capture. A microscopic model consists of a node-link system and can contain all necessary characteristics and parameters of the real infrastructure, such as switches, signals, speed and gradient profiles. The track layout can be defined with high accuracy. Macroscopic models represent data much more aggregated, stations can for example be defined by single nodes with attributes regarding the handling capacity. Links contain information on number of tracks, average speeds etc.

A full scale microscopic simulation model of a large network is data and work intensive but has the potential of giving results with high accuracy. Macroscopic models require considerable less data but cannot accurately capture interaction effects between trains which is important since the purpose of many simulations is to predict delays largely caused by train interactions (knock-on effects). However, macroscopic models can be used for long term traffic planning and strategic infrastructure planning for which high accuracy data is not required. Due to the difference in the size and type of data required, macroscopic simulation is generally considerable faster than microscopic simulation.

A trade off between microscopic and macroscopic models is to use a mesoscopic model. This will enable more detail and features in some areas, thus approaching a microscopic model, whereas a high level of aggregation is retained in other areas. It is for example possible to model station characteristics in more detailed and realistic way than in an exclusively macroscopic model (see e.g. Gille et al., 2008; Radtke, 2008).

Defining a microscopic rail network requires detailed data for the track layout with lengths, gradients and positions for switches where trains can change tracks. Furthermore information is needed for signal system objects, stop locations on stations and static speed profiles. Modeling signal system and interlocking characteristics requires effort and knowledge, to do it completely is often not feasible.

Train routes within stations, from entry to exit signals, model possible ways trains can utilize and pass a station. If information is available concerning route dependencies, e.g. a certain train route (movement) is not allowed while another route is set, they should be included in the model. How much detail is needed in an infrastructure model depends ultimately on the objectives and scale of the study.

3.2 Trains and timetables

Train run times, including acceleration and braking characteristics, are needed for designing timetables on a functioning infrastructure model. Typically a traction force diagram is defined for a locomotive or a train set. This data and the static and dynamic masses as well as resistance and adhesion coefficients describe the acceleration performance. In reality, also the braking performance of a train is partly defined by the mentioned parameters. The braking performance in traffic simulation models is often modeled with a constant braking rate.

Having created an infrastructure model and defined train performance characteristics, the timetable can be designed. A timetable defines origins and destinations for trains as well as the routes between them in a network, departure and arrival times at stations, scheduled stops and routes within stations. In this stage occupational conflicts between trains can be dealt with. The timetable is an operating plan that can be fully realized, assuming all conflicts are handled, if no delays occur. If the objective is to perform stochastic simulations, additional parameters are required.

3.3 Simulation with delays

The stochastic behavior of a railway system is emulated by introducing delay distributions and assigning them to trains at various locations. These are needed to emulate realistic exogenous events that may indirectly affect more trains and not just the first one, i.e. generate knock-on delays on other trains. Delay distributions reflect variations in passenger exchange times, run times, initiation in the network and so on. They can both be used to model regularly occurring smaller variations and systematic disturbances. Different kinds of delays appear with varying probabilities and magnitudes. Therefore, the probability of a delay affecting a train at a certain position is typically based on several distributions which reflect different possible events. As mentioned earlier, primary delays are introduced in the simulations in order to create secondary (knock-on) effects.

Other parameters are needed as well. This involves for example priority values for trains, which may be dynamic with respect to deviations from scheduled times, and look-ahead ranges. Alternative routes through stations are required to obtain flexibility in disturbed conditions. These type of parameters are provided so that the dispatching of trains is emulated as realistic as possible. The dispatching algorithm determines the order in which trains are processed at different locations. However, this is the very core of rail traffic simulations and also the part that is most difficult to model.

A timetable is normally run for multiple cycles to simulate a period of operation, e.g. a few months or a year. This is important for obtaining statistical stability in different measures of performance. This process is essentially a Monte Carlo simulation, a problem solving technique used to approximate the probability of certain outcomes by running multiple trial runs, called simulations, using random variables. Simulating a timetable just for a few cycles may not give representative results. How many cycles are needed depends on the size of the network, characteristics of delay distributions and which measures of performance are of interest. Evaluation of results can be done with respect to the quality of service measures discussed in section 1.3. In particular if a real network is studied, a calibration is most likely needed, meaning that adjustments in input data are made in order to produce results similar to observed data. A validation will further increase the confidence in a model. This process may take considerable time and effort in a large simulation project.

3.4 Discussion of the methodological problems

Depending on the purpose of the simulations and on the network model (real, imaginary or a combination) the delay modeling can vary. In some cases it can be sufficient to use some self-defined distributions, e.g. a low-high level approach. This can give insight on recovery performance, i.e. ability to reduce delays. For simulations based on existing networks historical delay data is commonly used. Three important delay types are usually considered prior to a simulation setup:

- Entry (initial) delays: Used for trains initiated at a network boundary or inside the network, i.e. entering into the network.
- Dwell delays: Mostly used for stochastic behavior in passenger exchange times or freight handling. Can also represent train or infrastructure malfunctions.
- Line delays: Run time extensions caused by infrastructure malfunctions, weather conditions, decreased vehicle performance, driver behavior.

Entry delays can often be extracted directly from historical delay data and are applied when trains are initiated in the network, either at a boundary location or inside. Dwell delays are difficult to compile into distributions based on normal delay registration data (Swedish conditions). This is both related to difficulties

separating primary and secondary delays and on the truncation of seconds in data. This is imprecise for estimations of dwell delays in most cases. Distributions compiled from manual measurements, carried out in previous projects, are used throughout this thesis (see Nelldal et al., 2008; Dagerholm, 2009). Distinction is made according to train category and station size with respect to passenger numbers.

Line delays are related to extended run times between stations. The secondary (knock-on) part is created in the simulations as reactions to other disturbances, i.e. conflicts with other trains. Realistic scenarios also demand a primary part for this type of delays. The normal train registration data records deviations from scheduled timetables at stations. Originally delay increases of five minutes or more between two adjacent stations or within stations were cause reported in Sweden. The Swedish Transport Administration has recently changed its policy and cause reports are now added from three minutes and up. If cause reported data is available, in theory a separation in primary and secondary causes is possible. However, the cause reporting is not consistent and delays growing slower than the limit between adjacent stations are not coded.

One way of creating smaller primary delays prior to a simulation is to make assumptions based on cause reported delays and then make estimations for the 1–5 (today 1–3) minute interval. Another approach is to disregard all cause reports and construct delay distributions directly from the basic registration data. Since this data includes all delays, without any information regarding the distribution of primary and secondary causes, the calculated distributions will need adjustments in order to reflect the portion of primary delays.

If the actual primary delays to be modeled are small or trains traverse a short distance in the network, run time extensions may possibly be excluded. It can then be sufficient to only model entry and dwell delays. However, in a large network with long distances the need for modeling run time extensions increases. Fig. 4 shows examples for two train groups on the Southern Main Line from simulations without applying run time extensions for any train. In particular, high-speed trains do not experience similar delay development as is the case for observed values from spring 2008.

The significant delay increase in observed data on the last section was caused by a speed limit due to long term tunnel maintenance. In this period, no additional allowance was introduced in the timetable. The freight train group has, relatively speaking, smaller differences considering that freight operations are more difficult to model than passenger operations. This is discussed in section 5.

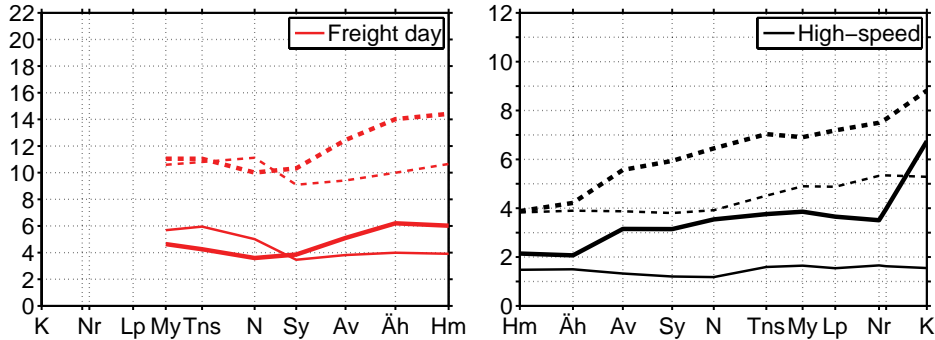


Figure 4: Simulated (thin) and observed delays (bold) in minutes, mean (solid) and standard deviation values (dashed). No primary run time extensions applied.

3.5 Limitations

Delay distributions used in the simulations of this thesis are designed with the purpose of avoiding large values that in reality would result in dispatching measures which are difficult to model in the simulation software. A typical real condition example is the case where trains on a double-track, where each track normally is operated unidirectionally, would be dispatched to use the track intended for traffic in the opposite direction. Although, this may work in many situations in the simulations, it has not been employed in this thesis in order to avoid so-called deadlocks. Limitations are also introduced for larger stations regarding possible routes between station entrance and exit to avoid deadlocks. When it comes to the Swedish signal system and the overlaying automatic train control (ATC/ATP), all features cannot be exactly modeled.

Shunting operation at stations is not modeled. This also applies for vehicle chaining and transfers between passenger trains. Although these characteristics can to some extent be considered in the simulations, they add more complexity and might not reflect actual conditions. Freight trains regularly depart ahead of schedule in Sweden, in the simulations presented in section 4 and partly in section 5 freight trains are either on time or delayed. Correspondingly this is also considered when performance measures are compiled from observed and simulated data. A modeling approach capturing this property is however presented in section 5.4.

The routing algorithm installed on RailSys is heuristic-based and not optimization-based (in mathematical terms) and there are thus certain limitations to the effectiveness of the dispatching measures. In this respect, routing is not designed to fix or optimize a perturbed scheduled timetable. In particular, this applies in cases where a large number of lines running in single-track or bidirectional operation are located within the area under review (RailSys, 2014).

In a synchronous simulation the operation is simulated in real time sequences and there is normally no parallel scheduling process during the simulation. Due to this, the situation becomes in a way less predictable in case of a delay. As a consequence, a synchronous simulation is prone to deadlocks on lines with bidirectional sections and on stations that have a complex layout (Pachl, 2011). If deadlocks mostly occur in certain locations and traffic situations, specific measures can be adopted in the infrastructure and routing parameters to counteract deadlocks. In congested situations, e.g. due to unscheduled operation or high stochastic delays in scheduled operation, the probability of deadlocks increases.

Additional measures may be necessary if the number of deadlocks is too high, i.e. simulation cycles cannot be completed due to situations the software cannot resolve. These can consist of both real deadlocks which, if they occurred in reality, would require reversing train movements or situations where limitations in the routing module create deadlocks. These problems occur mostly on networks with single-track lines and stations/junctions with complex routing. However, the number of deadlocks can usually be significantly reduced by making adjustments both in infrastructure and route settings. This means that the actual behavior in station routing is not always possible to model. Typical examples of deadlock situations are presented in fig. 5. Although the examples show single-track line layouts, similar situations can occur in complex stations or on multi-track lines where some or all tracks are operated bidirectionally.

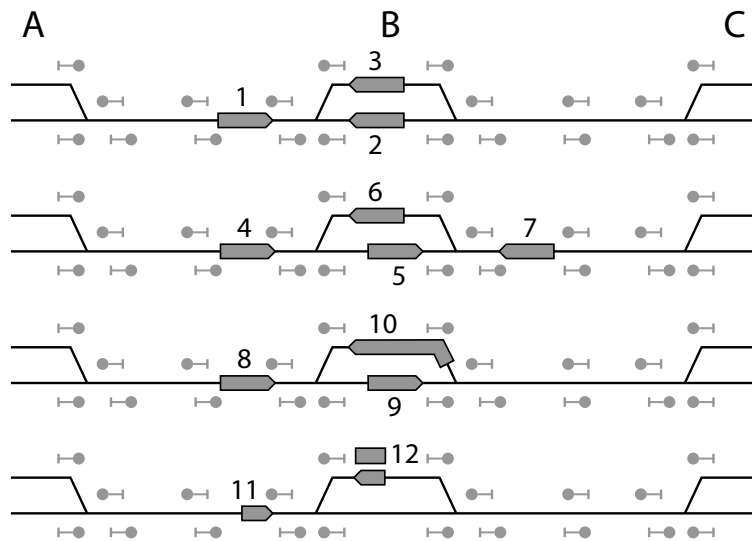


Figure 5: Deadlock situations on single-track lines.

The first example in fig. 5 shows train 1 in direction A to C and trains 2 and 3 in the opposite direction. To avoid a deadlock train 1 should have waited on station A or either one of the trains 2 and 3 should not have entered station B.

Similarly in the second example, either of the trains 4 and 7 should have waited on station A or C. In the third example train 10 is too long to clear the main track on station B, either train 8 or 10 should have waited on the preceding station assuming that train 10 can be handled there. The fourth scenario shows a case where both passenger trains 10 and 11 need to use the same track resource on station B in order to reach the platform.

In addition to the trains directly involved in a deadlock, there may also be other trains on surrounding stations that to some degree contribute to a resulting deadlock because they have been involved in earlier dispatching decisions. In real operations, a train with lower priority could be retained on an earlier station to avoid a congested situation further ahead. However, this is one of the significant differences between modeled dispatching in a simulation environment and real dispatching. Taking a decision for one train can affect multiple trains as time proceeds and depending on further decisions an overall solution is obtained. The handling of deadlocks when multiple timetables are simulated is described in section 7.7.

It should be clarified that the shortcomings with this type of simulation methodology compared to the real world are largely related to the look-ahead capabilities in space and time. In a real situation, decisions may be taken based on information of the type that would be hard to implement and use for predictions in a simulation environment. Canceling trains and performing turnarounds on stations other than the planned destination in order to maintain operations during or after a large disruption is not normally applicable in synchronous simulation softwares. This is the main reason for not introducing too large delays in the simulations.

4 Measures for improving on-time performance for high-speed passenger trains (Paper A)

The introduction of tilting trains ($v_{\max} = 200$ km/h) in Sweden 1990, enabled significant run time improvements on the Western Main Line Stockholm–Gothenburg and between Stockholm and Malmö, i.e. between Sweden’s capital and the second and third largest cities. Further infrastructure improvements have contributed in shortening running times. However, for the last ten years demand for passenger train services has increased considerably, while the freight transport has maintained high levels on those main lines. Several commuter train systems have expanded, both in frequency, distance and number of stops. The increased traffic also contributes to disturbance sensitivity. In this study simulation is used to evaluate how timetable adjustments for high-speed trains affect their on-time performance.

The Western Main Line links Stockholm with Gothenburg, the distance is 455 km. It is highly utilized on most sections. Commuter trains operate in both ends and several regional train systems use parts of the line. There are also InterCity-trains on the same relation, although some using a partly different route. Freight operations are significant, especially on the western part between Hallsberg and Gothenburg. The speed mix or difference between train average speeds is high on most parts. This limits capacity and adds to the disturbance sensitivity. In particular trains using higher speeds risk a relatively high delay increase when catching up slower trains or when affected by disturbances in general.

High-speed trains (X2000) have 15–17 departures on a normal weekday. Some seasonal variations exist and regarding different weekdays. On-time performance, a generally used measure describing outcome in rail networks, for high-speed trains has dropped under the last years as low as 60–70%. This has led to several studies, investigations and debates on how performance in the Swedish rail network can be improved.

4.1 Allowances and buffer times in timetables

Run time allowances are often included in operational timetables at some stage in the construction process. The purpose of allowances (supplements) is counteract so that smaller disturbances do not necessarily grow into larger delays which may not be completely recovered from during the train run. An allowance can serve a local purpose, e.g. before a junction station, so that traffic from two lines can be merged with better precision. It may also be a more general run time allowance that is distributed more or less evenly along a train journey. Some allowance can also be added at stops, i.e. trains may be able to do a shorter stop than what is scheduled, and this can facilitate delay reduction (see e.g. Rudolph, 2003).

Basically an allowance adds time to a certain minimum run time. This means that trains would not normally have to run with maximum speed at every moment since the allowance makes it possible to negotiate the schedule with a slightly lower speed. In the case with minimum run time, a speed curve considering a certain acceleration and braking depending on the train type defines where the train will be and what speed it will have at every moment on its journey. Adding allowance could mean that acceleration and braking ends and starts from a lower speed than defined by the minimum time diagram. The allowance can also be handled by making a slower acceleration and longer braking. Allowance does not necessarily have to be consumed, as long as a train does not have to make a scheduled stop and adjust a departure time or interfere with other trains the allowance can be saved for future use.

Buffer time is the temporal space between one train and other preceding and following trains. For example, if one train is crossing over main tracks in a station and this is followed by another train using one of the main tracks then these two movements must be separated in time. This type of separation is needed everywhere where two consecutive trains will use, partly or in full, the same track resource. Settings short times will allow more trains to be scheduled, but trains can easily start interacting and spread delays in case of an initial disturbance. Setting longer times, i.e. ensuring there is some extra buffer, creates more space in the timetable at the expense of the number of trains that can be operated.

4.2 Case study

The influence of changing the allowance and increasing the buffer times is studied by doing operational simulations for the high-speed passenger services between Stockholm and Gothenburg. This study is based on an infrastructure model used in a previous study where the focus was to make cause specific reductions in delay distributions and evaluate the impact on on-time performance (Nelldal et al., 2008). Specifically, the aim was to reduce delays caused by infrastructure, vehicle and operator related problems. Some rebuilds and additions are made in the model, including balises and speed settings. Delay input (distributions) is reused from this earlier study with additions from another simulation study done on the same line (Lindfeldt and Sipilä, 2009).

Parts of connecting lines are added in order to capture the potential conflicts occurring at some stations due to in- and outflow on the main line. The influence can be even more substantial considering that some are single-track lines. Whether this approach is necessary or not depends for example on the frequency of crossing conflicts and limitations on available station tracks. The time interval between outgoing and incoming trains is also modeled in a more realistic way.

The focus is mainly on adjusting timetables and evaluating outcome for high-speed trains. Applied delay distributions and dispatching parameters are not changed in the scenarios. High-speed train timetables between Stockholm and Gothenburg

have a general 3% driver allowance added to the minimum run times. Node allowances are used to maintain a certain precision even if small disturbances occur, e.g. extended dwell times, variation of speed due to restrictive signal aspects, etc. These are usually applied in the range of 4 min per 200 km, resulting in 7–8 min between Stockholm and Gothenburg. Rounding up to full minutes adds to the total allowance.

The modeled timetable is considered to represent a normal weekday, Thursday 29th January 2009, full traffic is assumed. Passenger train timetables are usually similar comparing different working days but freight train operations can have significant variations. Three alternative simulation scenarios for high-speed trains are defined, aside from the reference case:

- Reference, scheduled timetable year 2009
- Decreased allowance with 4 minutes
- Increased allowance with 4 minutes
- Increased buffer time separating high-speed and other trains

Decreased and increased allowances are applied in the way that half of the timetable adjustments are made on section Stockholm–Hallsberg and the other half on section Hallsberg–Gothenburg. Buffer times are increased so that high-speed and other trains both are separated by at least 5 minutes on lines and at stations with overlapping routes, this refers to timetable time. If block occupation time is compared the buffer time decreases. A buffer time of 4 minutes is allowed in situations where, for example, a freight train is scheduled to a siding followed by a high-speed train on a main track. Timetable for year 2009 had buffer times down to 3 minutes regarding high-speed trains at some stations. Buffer times on the bottle-neck south of Stockholm Central Station are not considered, these are 2–3 minutes.

Available data regarding real outcome is from August 2006 to January 2007. Simulation of the reference timetable is compared to this data, some adjustments are made to delay distributions and additional trains due to the extended network. Adjusting the timetable is made with the aim of minimizing number and magnitude of changes for other trains. However, it is not possible to avoid disrupting regular interval passenger train timetables and in some cases completely reschedule freight trains.

4.3 Results

On-time performance diagrams are presented on aggregated level for all high-speed trains with respect to the different timetable cases. On-time performance for the reference case increases close to the end station in direction from Stockholm to Gothenburg. This is due to allowances placed on this section. A similar allowance exists on the bottle-neck between Stockholm Central and Stockholm

South Stations, a double-track section of around 2 km. There are differences between the simulated reference timetable and the comparison timetable that was used during the period from which registration data is obtained, both for high-speed trains and other trains. However, the purpose of this comparison is to check the behavioral trend of the simulated on-time performance.

Some sections are problematic, especially if compared to the relative distance. This can be explained by both dense traffic with limited overtaking possibilities and also by tight schedules with regard to minimum running time. Fig. 6 shows that on-time performance at terminus decreases with approximately 10 percentage points in both directions with decreased allowances.

If allowances are increased improvements are higher for trains from Gothenburg to Stockholm compared to the other direction, 10 and 5 percentage points. This difference in directions gets clearer comparing the case with increased buffer times. Improvements are also in this case considerably higher in northbound direction, for southbound trains the difference is marginal. One reason for this is that in northbound direction, more timetable changes are needed to satisfy the minimum buffer time requirement. Thus, there may be a higher potential for improvement.

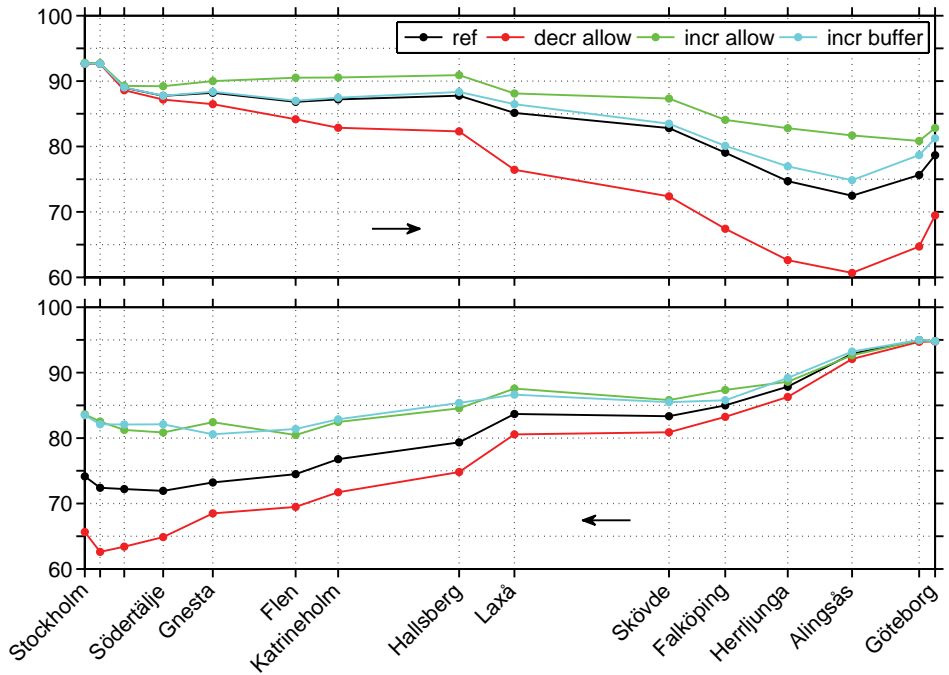


Figure 6: Simulated on-time performance (%) for reference and alternative timetable cases. Southbound (upper), northbound (lower). Trains with delays less than or equal to 5 minutes are considered on time.

Simulated mean and standard deviation for arrival values are generally lower in Stockholm compared to Gothenburg. A late high-speed train faces the risk of ending up behind a commuter train on the section Alingsås–Gothenburg (45 km). These trains make several stops and overtaking possibilities are limited. In the simulations, no overtakings between passenger trains are allowed on this section. The situation for northbound trains close to Stockholm is not as sensitive compared to southbound trains close to Gothenburg. The commuter trains have separate tracks, the speed mix is lower and most high-speed and regional trains have the same number of stops.

Fig. 7 shows the difference in mean and standard deviation arrival values between reference and alternative timetables for each high-speed train. Positive values indicate improvements. Mean arrival values for the reference case are typically in the range of 2–5 min in Stockholm and 3–6 min in Gothenburg. Corresponding standard deviations are in the range of 6–10 min in Stockholm and 10–13 min in Gothenburg. Both directions have 3–4 trains that fall outside these intervals, southbound direction has the highest values. Fig. 7 can be compared with fig. 6, generally the case with increased allowance shows the best improvement for standard deviation, the difference for mean values is not as clear.

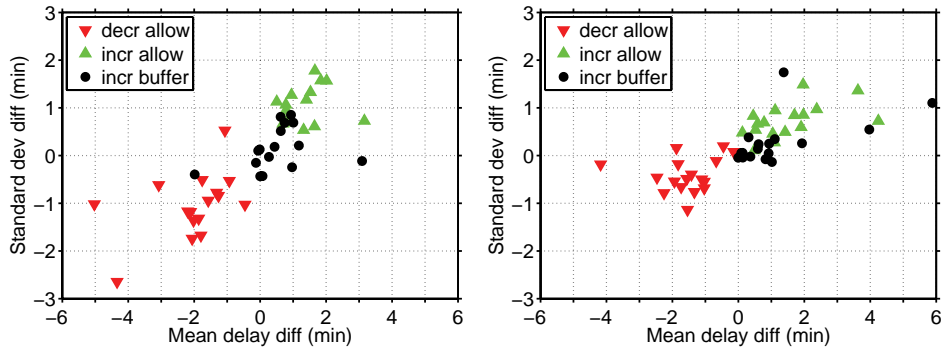


Figure 7: Difference in mean and standard deviation for arrival values between reference and alternative timetables, southbound to Gothenburg (left) and northbound to Stockholm (right).

Changing from aggregated to individual level helps in finding specific trains and sections that may have a relatively big influence on aggregated on-time performance levels. Fig. 8 shows the situation for some trains, both reference and the case with increased buffer times. Two of the trains have significant drops in on-time performance between Gnesta and Södertälje. In these cases, time interval between passing high-speed train and departing commuter train is short. The common section is between Gnesta and Järna (17–18 km). For a late high-speed train there is a risk of getting behind the commuter train, instead of the other way around. There is no overtaking possibility in the simulations and it is also hard to accomplish in reality.

This scenario with small buffer times (3–4 min) at Gnesta occurs for six high-speed trains from Gothenburg in the reference timetable. In the case with increased buffer times, the commuter trains in question are pushed. The effect of this is shown in fig. 8 and the level of on-time performance can be maintained to the end station without the reduction occurring in the reference timetable. The other exemplified trains show improvements in on-time performance as well. Considering all trains, the majority get improvements. Trains that already have a high on-time performance get only marginal or no changes, indicating that these trains already have a good slot regarding buffer times.

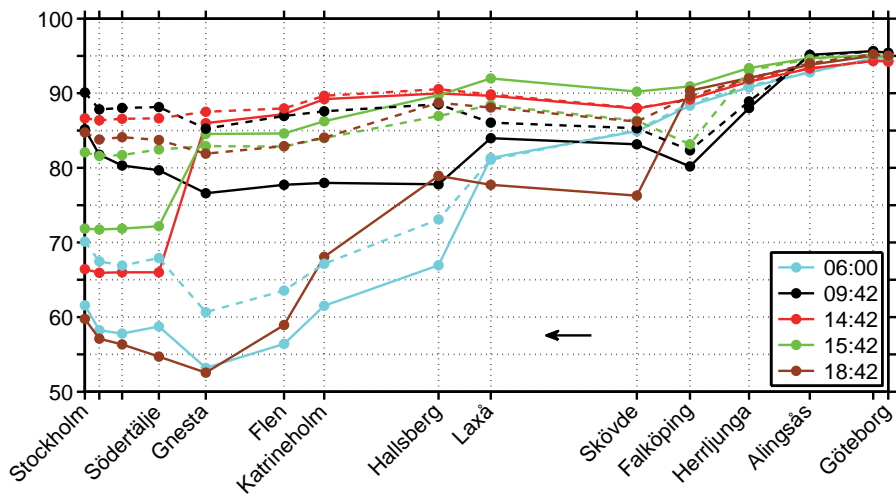


Figure 8: Simulated on-time performance (%) for selected northbound trains, reference timetable (solid) and case with increased buffer times (dashed). Trains with delays less than or equal to five minutes are considered on time. Legend shows departure times from Gothenburg.

4.4 Conclusions

Simulations show that if allowance is decreased with 4 minutes, the arrival on-time performance drops from 79 to 70% in Gothenburg and 74 to 65% in Stockholm. The opposite scenario, with increased allowance, can improve on-time performance with 5–10 percentage points depending on direction. The improvement on aggregated level is larger for northbound trains. However, on-time performance is higher for southbound trains in the reference case.

If the buffer times are increased for high-speed trains, on-time performance for arrivals in Gothenburg improves from 79 to 83% and in Stockholm from 74 to 84%. Northbound direction shows a larger improvement which can partly be explained by the directional difference in the number and magnitude of timetable changes. Obtaining the defined buffer time limits required more adjustments

for northbound trains in the reference timetable, thus implying that the relative improvement could be more pronounced.

Moving from an aggregated to individual level shows that some trains suffer from large drops in on-time performance or increased mean delay on some sections. This can both be explained by locally tight situations with other trains, e.g. slower commuter trains coming in front of delayed high-speed trains, and also by the difference in real and modeled dispatching. Increasing buffer times between passing high-speed and departing commuter trains in Gnesta give clear improvements in the simulations. This strategy has also been adapted in reality and used in year 2010 timetable. Some aspects concerning on-time performance on this section and crossing train routes in Gnesta are discussed in Anand and Anayi (2009).

Increased buffer times between high-speed and other trains is a strategy that could be used in gaining improved on-time performance. However, the space for other trains reduces meaning that some trains may not fit in the considered time space. If buffer times for other trains are squeezed making them more sensitive to delays with propagation effects this can eventually also affect the high-speed trains negatively. Further investigations of this can contribute in finding a realistic trade-off that can be used and give satisfactory results for the whole timetable.

On-time performance is checked on an aggregated level for other passenger trains and freight trains on the same evaluation stations as for high-speed trains. The amount of trains varies from station to station since no chaining is done. Results of this show only marginal differences for passenger trains, on-time performance for freight trains have the same pattern on most stations. Improvements are achieved in the case of increased buffer times at some locations. A more detailed description of the results from this work is presented in Sipilä (2010).

5 Method for calibrating primary run time delays (Paper B and C)

One of the difficulties in preparing input to simulations is to design realistic run time extensions, i.e. primary delays that are applied on trains between stations. The importance of using this delay type is emphasized in previous simulation studies in Sweden (Lindfeldt and Sipilä, 2009; Nelldal et al., 2008). If only dwell extensions and initial/entry delays are modeled, the delay increase (punctuality decrease) along a studied line typically does not reflect the historical data it is compared with. This depends of course on other aspects as well, e.g. traffic density, available allowance etc. In this section, a method for estimating primary running time extensions from historical data is presented.

5.1 Background

Run time extensions are used for modeling variations in travel times from one station to another. These extensions can have several causes, e.g. decreased acceleration performance caused by weather related or vehicle problems, line speed restrictions, signaling failures etc. Freight train configurations not matching their planned schedule, e.g. higher train weight and lower top speed, can also cause delays. Knock-on delays alone, are not causing all variation between observed and scheduled run times.

Delay statistics can be divided into two parts. General statistics show deviation relative a planned schedule, mostly at stations, and have a resolution of one minute. This data does not indicate on delay causes. The other part consists of delays with cause reports. These are based on limits where a train's run time between two stations is compared to the scheduled run time. A cause report is added, if a delay increase exceeds or is equal to a specified limit. This data can later be sorted into primary and secondary (knock-on) events.

In an ideal situation, the cause reported data would capture all delay events (small or large) and the construction of distributions for run time extensions would be easy. However, the time limit used means that smaller delays are not reported. This also means that a delay that grows over several stations but do not reach up to the time limit between registration points is not reported, although the final delay may be large. A delay increase has to be at least five minutes in order to get a cause report (year 2008).

Cause reported delays are not used in this study. Instead delay distributions are estimated from regular statistics based on train registrations. The idea is to change the initially obtained distributions, since these contain a mix of both primary and knock-on delays and apply this in a simulation model on parts of the Southern Main Line. Evaluations are done for the section Katrineholm–Hässleholm, although the simulated network is extended and includes some of the connecting lines.

5.2 Method

The data used for estimating run time extensions is based on general delay statistics (registration data) from January to June 2008. A regular Thursday in January 2010 is chosen as reference for the timetable. Additionally trains from the previous and following night are defined making the simulated timetable 32 hours in total. Connecting lines are included to the first station with the aim of capturing crossing effects on the main line and possible time differences between outbound and inbound train movements.

The main line is divided into sections of 35–50 km in such a way that most trains using a certain section traverse from station A to F or F to A (fig. 9). Delay statistics are obtained for junctions and turnaround stations defining the sections and network boundary stations. For example, in fig. 9 there is no data available for stations B to E. Most trains are identifiable with this approach, although a few trains cannot be entirely tracked.

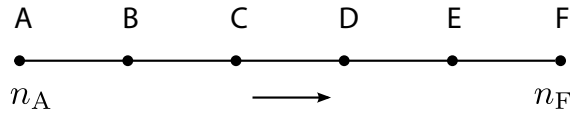


Figure 9: Definition of line sectioning with registered delays n_A and n_F .

Saturdays, Sundays and holidays are filtered out from the data. Due to a major disruption on one of the sections, causing significant delays, seven regular working days are excluded. Data is sorted on stations and additional processing is needed to identify paths for individual trains. Only data for trains (train numbers) running on minimum 20% of the total evaluated days are used for making distributions.

For all qualified trains the difference between n_A and n_F is considered. Positive values indicate delays and negative values mean that trains are ahead of their schedule. However, this is not modeled in the simulations at this points and therefore negative registration values default to zeros. Four different cases are possible for the difference value (t) and they can be described by

$$n_A \geq 0 \rightarrow t = \begin{cases} n_F - n_A & \text{if } n_A \leq n_F \\ 0 & \text{if } n_A > n_F \end{cases} \quad n_A < 0 \rightarrow t = \max\{0, n_F\}$$

Base distributions with t -values are compiled for different train groups, sections and directions. Fig. 10 gives an example on observed and adapted run time deviations. In this case pure deviations are considered and not the conditions used for trains ahead of schedule discussed above. Fig. 10 clearly illustrates the varying freight train performance. Passenger train groups are mostly well defined but freight trains show significant variation in running times and train configurations.

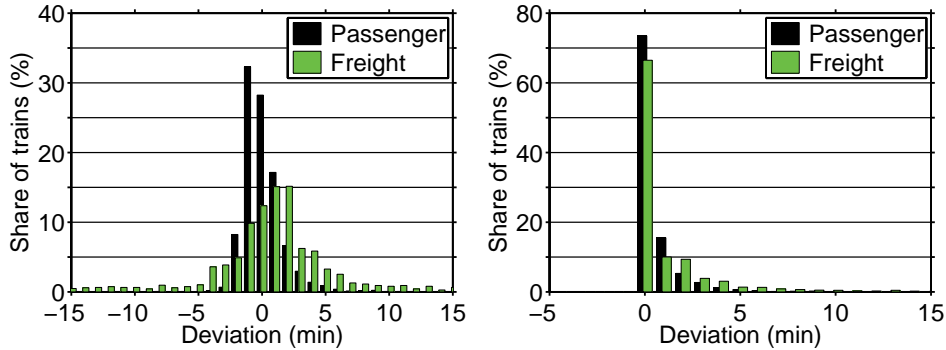


Figure 10: Deviation from scheduled run time (minutes). Southbound trains Mjölby–Tranås. Positive values indicate delay increase. Observed distribution (left) and partly adapted distribution for simulations (right).

Freight train grouping is based on running distance and further split into day and night groups. Distributions are reduced by keeping a certain percentage of values for every minute level, i.e. the remaining number of values are regarded as primary run time extensions. Number of registrations are kept constant meaning that removed registrations with an actual delay are considered as zero values.

Reductions are applied in four levels (20, 40, 60 and 80%) and varied separately for freight and passenger trains, resulting in 16 simulation setups. The results are later compared to real outcome and evaluated by using the root mean square error measure (RMSE) for mean and standard deviation values.

Dwell and entry delays are handled as described earlier. Freight trains are matched on train numbers or time and grouped according to mean and standard deviation values in the simulations. Scripts are needed to simplify construction of input delay files used by the simulation software. An additional benefit is that random seed can be controlled meaning that other applied non-varied delays are identical in the simulations.

Run time extensions have not been applied on each subsection between two adjacent stations. Instead the network is divided into larger subsections, where the boundary stations typically are junctions or stations with regular turnarounds. The principles used are similar to the ones presented in e.g. Lindfeldt and Sipilä (2009); Sipilä (2010). Distances for subsections used in simulations on the Southern Main Line are in the range of 35–50 km with 2–6 intermediate stations.

Run time extensions have typically been applied in the middle of subsections. This assumption is checked by simulating cases where different approaches are used for positioning the extensions. Three additional cases are modeled and positions are set to the first or last section combined with randomly assigning a section (uniform distribution). Delay values are equal in each case.

Fig. 11 shows mean and standard deviation values for southbound high-speed and northbound daytime freight trains. The difference is marginal in the first case, although sections with high run time extensions are clearly distinguishable. The difference between delay development for the freight group is not large, in a relative sense. Levels for run time extensions are higher compared to the high-speed trains.

Due to a lower dispatching priority and speed freight trains frequently use sidings, hence allowing faster passenger trains to pass. This procedure, if not pre-scheduled, results in knock-on delays and adds to the total delay development. Considering these four delay positioning cases, using randomization seems to be a good alternative and most realistic considering that it models the stochastic behavior of actual delay positions.

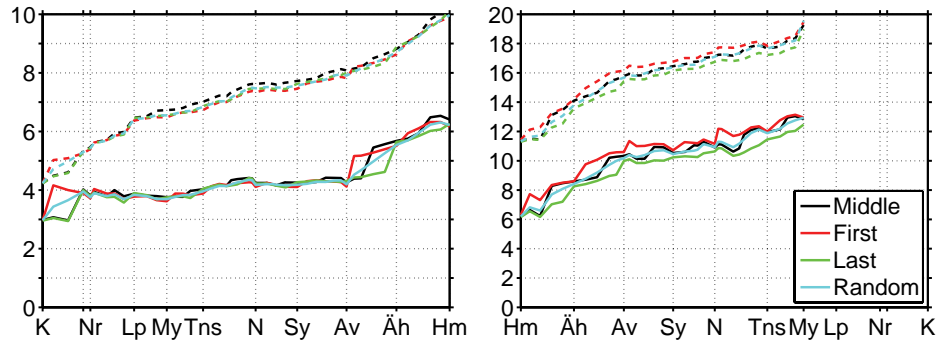


Figure 11: Simulated outcome in minutes with different schemes for positioning run time extensions on sections. Southbound high-speed trains (left) and northbound daytime freight trains (right). Mean (solid) and standard deviation values (dashed).

5.3 Results

Delay cases are evaluated by comparing simulated and real outcome and using RMSE as measure of performance (MOP). Comparisons are made for six representative train groups at stations for which real delay data is available. Each group uses at least three measurement stations. Fig. 12 shows results for the applied combinations of run time extensions in southbound direction. Freight train groups have in many cases generally higher values than passenger trains, which can partly be explained by greater differences between planned schedules and actual operation. The character of the originally defined distributions are such that reductions give a higher impact compared to passenger trains.

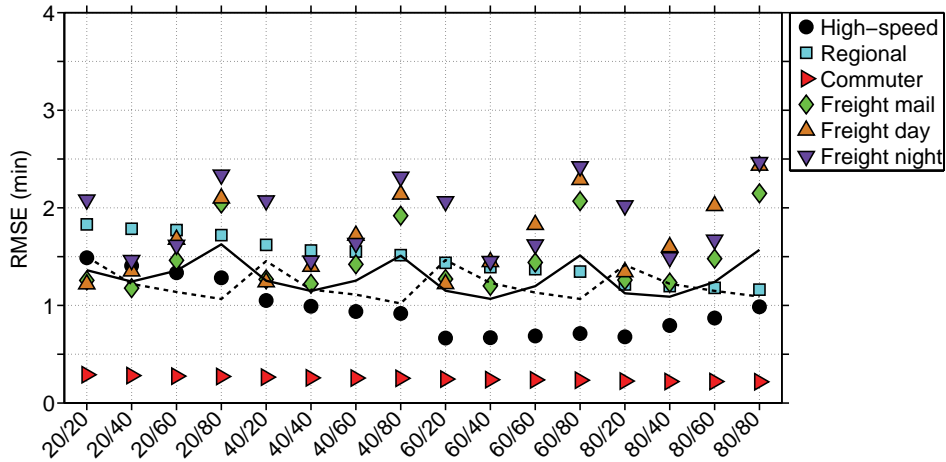


Figure 12: RMSE points for mean values. Solid line shows mean value for RMSE points and dashed line for RMSE standard deviation points. Reduction percentages on x-axis, passenger/freight.

The commuter train group (Norrköping–Linköping–Mjölby) has almost no variation comparing the different cases. The probability of getting a primary run time extension is low and varying reduction levels have little influence. This can also be an indication that allowances make up possible delays efficiently. All of the passenger train groups have little variation when the reduction level is changed for freight trains. These groups are in most cases given a higher dispatching priority than freight trains.

RMSE values for northbound direction show a higher degree of variance for freight trains compared to the other direction, the original unreduced distributions show large deviations in run times. Passenger train variation is relatively small. The reductions used are always same in both directions meaning that a recommended case should simultaneously give acceptable results in both directions. Alternatively, if one assumes a low level of interaction between directions, two different cases can be chosen. However, this should be simulated and checked.

In fig. 12 freight trains have the lowest RMSE values if original distributions are reduced to 40%. Best fit for passenger trains is on the 60% level, especially high-speed trains are assigned more weight than the other groups. However, the freight train groups have high spread and changes from one delay level to another do not correlate well, i.e. there is no single delay level where all freight groups have minimum values simultaneously.

Fig. 12 gives only a goodness-of-fit measure and do not indicate on actual delays. Plotting delay development throughout the studied line is descriptive since possible problem areas can be identified. Fig. 13 shows mean and standard deviation values for delay case 60/40. It is clearly observable that agreement for freight trains is worse than for passenger trains. There is usually no good simultaneous agreement between day and night freight train groups and between mean and standard deviation values.

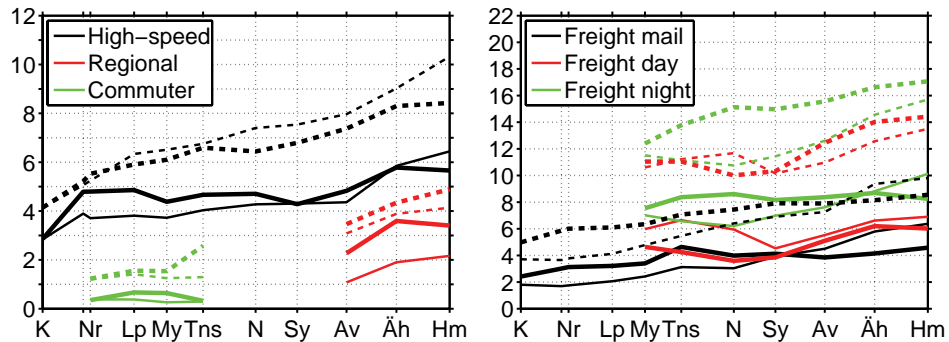


Figure 13: Simulated (thin) and observed outcome (bold) in minutes, mean (solid) and standard deviation (dashed) for 60/40 case. Passenger trains (left) and freight trains (right).

Comparisons between observed and simulated data can also be made by checking exit delays given an entry delay interval. This can for example give information on the possibility of reducing an entry delay and give some insight on the breakpoint, i.e. if an entry delay is in this range or higher the possibility of significantly reducing this delay before exit is small.

Table 2 shows the number of registrations used for the intervals in fig. 14. Simulation values represent case 60/40 according to fig. 12. In the lower start delay intervals, many of the trains are able to remain in the same interval on exit, alternatively reduce their delay completely. The higher start delay intervals show a larger spread on exit, the relatively small number of registrations adds to this.

Considering freight trains, they are both prone to delay increase and in the same time have the possibility of reducing delays considerably. The first issue is due to their dispatching priority compared to passenger trains, this holds especially in the simulations. Many of the so-called timetable technical stops for freight trains can be cancelled or at least reduced in time depending on the operative scenario. This makes it possible to catch up on delays.

Table 2: Number of observations in start delay intervals used in fig. 14

Delay	Southbound high-speed trains					Northbound freight trains				
	0-2	2-4	4-6	6-8	8-10	0-12	12-24	24-36	36-48	48-60
Reg	715	290	143	81	46	1150	125	40	33	22
Sim	1368	522	245	145	97	4071	449	89	72	55

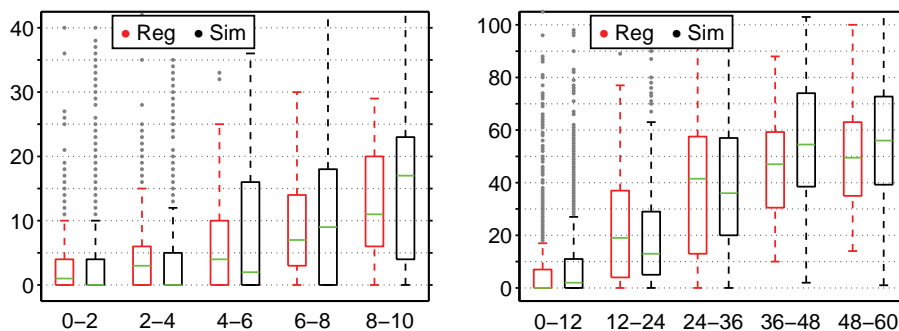


Figure 14: Observed (Reg) and simulated (Sim) outcome in minutes for trains depending on start delay intervals (x-axis). Southbound high-speed trains Katrineholm-Hässleholm (left) and northbound freight trains Hässleholm-Mjölby (right).

5.4 Modeling freight trains ahead of schedule

Passenger trains are rarely ahead of their schedule in real operations, when it happens it is usually in the range of a few minutes. However, freight trains can have a high spread in their registrations. It is not uncommon with freight trains ahead of schedule. This may be attributed to the origin station or it may evolve during the train run. Reasons for a train, that starts late or on time, to get ahead of its schedule further on are for example:

- The schedule is planned for trains with characteristics other than is operated
- Commercial stops are realized in shorter time than planned
- Non-commercial stops induced by meets or take-overs are canceled

The effect of trains getting ahead of their schedule can to some degree be captured in the simulation by setting different parameters for scheduled stops. However, the trains can only be initiated either on time or delayed. Fig. 15 shows a departure distribution for freight trains. Almost 50% of the departures happen ahead of schedule. If no further action is taken in a simulation, the real departure distribution is transformed to the adjusted distribution, i.e. trains ahead of schedule are treated as if on time. This means however that the variance in freight operations is not captured entirely.

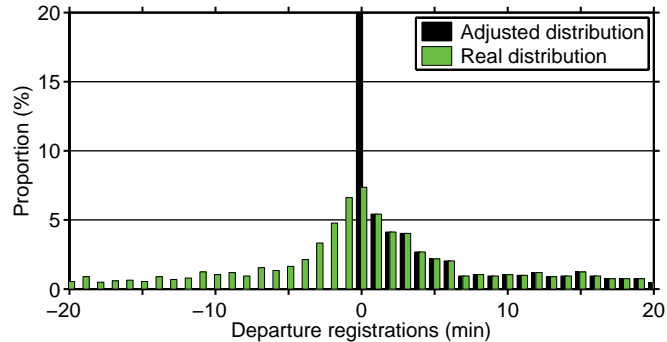


Figure 15: Example of departure distribution for freight trains. Positive values indicate delays. Zero level is 45% for the adjusted distribution.

A way to include the real distribution is to adjust freight train schedules by shifting them in time so that the zero point in the distribution coincides with the scheduled departure time. How large time shifts should be used depends on the where the left tail of a distribution is truncated. As mentioned earlier, in using observed data train groups can be formed, e.g. by using cluster analysis, with respect to their respective mean and deviation values. Some trains have a better precision than others in the registration data and this is better described in a simulation if trains are split into subgroups when distributions are assigned.

To investigate how these proposed measures affect the simulation results compared with the case where freight trains are not initiated before schedule, an operational simulation is performed based on the findings in section 5.3. The difference is that all freight trains are shifted 60 minutes ahead of their respective schedules and the initial distributions are changed accordingly. Parameters for dispatching priorities are defined so that trains that are close to being on time have a higher priority. Additionally, late trains have a higher priority than trains ahead of schedule. No changes are made on passenger trains or other trains not classified as freight trains.

Simulation results show no significant difference for evaluated passenger train groups compared with the original case. Regarding freight train groups, they will in general have lower average delay values and higher standard deviation values compared with fig. 13 and corresponding data in the northbound direction. The difference between observed and simulated values is mostly higher in this case. Although it could be anticipated that the passenger train groups would be more affected due to the higher operational variance for freight trains, the priority based dispatching is able to counteract for this variance and leave passenger trains basically unaffected.

Instead of looking at e.g. average delays and on-time performance, it is also interesting to study the development along the line or network given an initial interval representing schedule deviation. Fig. 16 shows observed and simulated

northbound cumulative distributions for three start deviation intervals and three stations, the principle is similar to that in fig. 14. The three deviation intervals are defined by 30 to 10 minutes ahead of schedule, 10 minutes ahead to 10 minutes behind schedule and 10 to 30 minutes behind schedule.

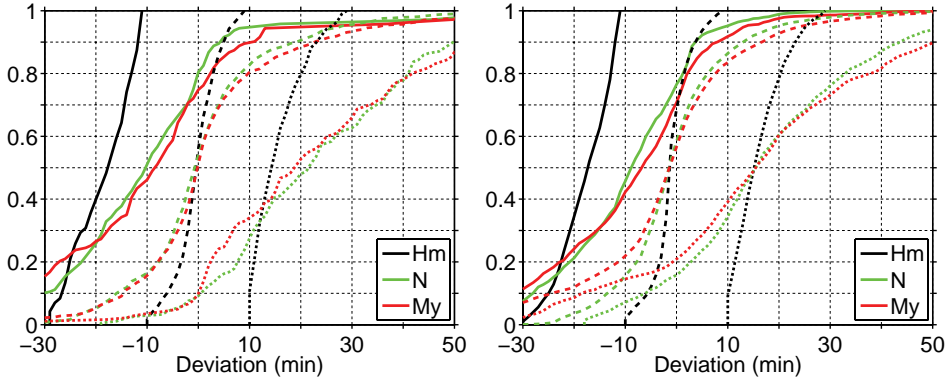


Figure 16: Distributions showing observed (left) and simulated (right) deviation development for northbound freight trains. Start intervals are defined by distribution in Hässleholm (Hm). Positive values indicate delays.

The middle interval (± 10) represents around 75% of the total number of trains in all three intervals. The remaining 25% splits approximately in equal halves on the other two intervals. One can observe, considering the middle interval, that 60% of the original trains remain in this interval traversing from Hässleholm to Mjölby (273 km). Remaining trains are more than 10 minutes before or after schedule. There is a good agreement between the observed and simulated outcome, this also holds for the southbound direction. If the simulation captures the handling of freight operations poorly, the cumulative distributions would give an indication of this and show less agreement.

5.5 Conclusions

Results show that a reasonably good fit is possible to obtain using deviations from scheduled run time as the basis for estimating primary run time extensions. The subdivision used is relatively rough, a finer breakdown can give a better estimation. However, it must be considered that the scheduled timetable is rounded to full minutes and if the distance considered is small this can contribute with a relatively large error. Using data for every station is valuable also for analysis of real outcome and this will add knowledge that is useful in setting up simulations.

A general problem with the observed data (registrations) is that arrival times are only reported according to scheduled stops. This means that stops made on other stations cannot be identified in a straightforward way. Comparing run times and checking for changes in train sequence is needed in these cases. Track circuit data

can also provide a solution to this, but processing this for many stations on a large network can be challenging. Since passenger trains rarely make this type of unscheduled stops on double-track lines, it mainly concerns freight trains.

Passenger train operations show a better fit than freight trains. Difficulties in modeling freight operations are for example length and weight variations, top speed limitations and stops including shunting. Dispatching priorities can also vary for freight trains in real operations, this is not considered in the simulations. Modeling freight trains ahead of schedule to better capture the actual variance in freight operations showed only marginal effects on the evaluated passenger train groups compared with the original case. Conclusions from this study is that if passenger trains are in focus and they have a higher dispatching priority than freight trains, it would be sufficient not to consider the full variance in freight operations. If the focus of a simulation study is on both train categories or mostly on freight trains, then this methodology can be of use.

Exact train configurations were not known meaning that assumptions regarding weight, length and maximum speed were necessary for some of the simulated freight trains. Train configuration parameters are included when paths (slots) are requested. In addition to engine data as well as weight and length data, this means e.g. that some of the granted paths are setup for trains with maximum speed 70 km/h, whereas others are setup for 100 km/h or higher speeds. The operated train configurations may differ from the ones stated in the train path information. Operating a significantly heavier or lighter train, with respect to the train path definition, affects acceleration performance and to some degree braking performance.

It should also be pointed out that the simulated timetable and observed data stems from different time periods. Although many of the train slots are close to identical, there are differences and these mostly concern the freight trains. Using observed data from a longer time period and comparing outcome from different timetable periods is recommended.

The Swedish Rail Administration, later incorporated in the Swedish Transport Administration, conducted a one week trial in March 2009. The purpose was to observe differences in delay outcome between normal operations and operations where freight trains are not allowed to depart before schedule (Banverket, 2009). Conclusions made from the study are consistent with the simulation results, namely that passenger train performance is only marginally affected and that the average delay level increases (on-time performance decreases) for freight trains if clearing before schedule is not allowed. Another conclusion is that freight trains will mostly disturb other freight trains, this follows from the principle that passenger trains to a great extent have a higher dispatching priority.

6 Calculation of run times for the Green Train (Paper D)

In the Green Train research and development program (in Swedish *Gröna tåget*) a new generation of high-speed train concepts are studied considering a system perspective. The objective has been to develop a concept proposal for a new, attractive high speed train adapted to Nordic conditions that is flexible for several different tasks on the railway and interoperable in the Scandinavian countries. This includes both technical, safety and environmental issues combined with market and economical analysis. Relations between vehicle configuration and infrastructure as well as operations and capacity are of special interest. Innovative solutions can improve both performance and economy which in turn contributes to better socio-economic conditions for rail traffic (Fröidh, 2012).

6.1 Background

Market and economy is a subproject within the Green Train program and deals with traffic patterns, run times and other topics which are of importance in creating an attractive product for travelers. This involves for example train interior design with comparisons between seat layouts, pricing and load factors. People's dispositions for choosing the train instead of other alternative means of travel are, in addition to comfort issues, also highly dependent on travel times and in a wider perspective on reliability, i.e. a high share of the trains should be on time, which is a quality issue.

One key factor in attracting passengers is shorter run times. Fast trains have a comparative advantage over car and air travel on medium distances, i.e. 300–600 km. Travel times can be significantly lower than for cars but also compared to air travel if airport procedures and journeys to and from airports are considered.

High train speeds usually require new infrastructure with large curve radii. This makes it possible to separate high-speed trains from those with lower speeds and improves the capacity situation. There is also a need to control new bottleneck situations which are likely to arise in some areas and in junctions where new and old lines connect. Another possibility is to increase speeds on existing lines. This can be achieved by using modern trains with track friendly running gear and adjusting some track parameters. Updates in other infrastructure are also necessary at some locations.

As mentioned one of the main limitations in achieving higher speeds on existing lines is curve characteristics. Most of the main lines in Sweden have some longer sections with relatively straight track geometry. Adjusting cant in curves can improve static speed profiles for certain trains. One main focus in this subproject is therefore to analyze curve radii data on six dedicated lines and use different cant and cant deficiency set values to calculate new static speed profiles.

Other parameters affecting run times are top speed, acceleration and braking performance of the train itself. This relates to tractive effort, power output, train resistance etc. An obvious impact on total traveling times comes from stopping patterns, i.e. number of station stops. The main objective is to vary some of the characteristics affecting run times. In this study several new static speed profiles according to cant and cant deficiency parameters are considered. In addition different start acceleration and power output settings are evaluated. Calculations are done for several different routes and stopping patterns.

6.2 Calculation of maximal curve speed

The track geometry is a crucial factor for top speed and, by extension, run times. Sharp curves, i.e. small curve radii, imply limitations on maximum permitted speeds. Using trains with carbody tilting and bogies with good running characteristics can, to some extent, compensate for sharp curves. New speed profiles are based on existing line speeds, circular curve radius data and combinations of cant (D) and cant deficiency (I). Applied cant, also called super elevation, is the amount by which one running rail is raised above the other running rail. Cant deficiency is a measure of the resulting lateral acceleration (centrifugal force) in curves and is measured in the track plane (Andersson et al., 2013a).

Circular curves have characteristics which explain some of the steps used in the calculations. The curvature is constant, i.e. the radius is constant. For a constant speed circular curves give constant quasi-static acceleration, which justifies a constant cant value. Fig. 17 illustrates some of the definitions.

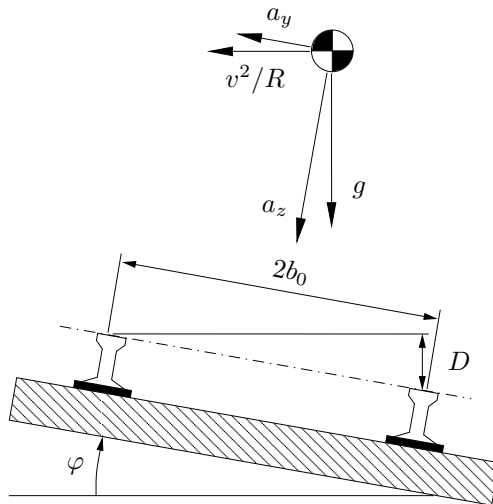


Figure 17: Horizontal (v^2/R) and vertical acceleration (g). Acceleration parallel (a_y) and perpendicular to track plane (a_z).

From fig. 17 follows that for a given speed (v), curve radius (R) and cant angle (φ) lateral acceleration (a_y) is calculated by eqn. 1. If φ is small the simplification in the second expression can be used.

$$a_y = \frac{v^2}{R} \cdot \cos \varphi - g \cdot \sin \varphi \approx \frac{v^2}{R} - g \cdot \frac{D}{2b_0} \quad (1)$$

A cant value that exactly balances the quasi-static lateral acceleration ($a_y = 0$) for a given speed and radius is called equilibrium cant (D_{EQ}) or theoretical cant. This can be estimated from

$$D_{\text{EQ}} \approx \frac{2b_0}{g} \cdot \frac{v^2}{R} = \left\{ v : \text{m/s} \rightarrow \text{km/h} \right\} = C \frac{v^2}{R} \quad (2)$$

Assuming standard gauge gives that $2b_0 = 1500$ mm. Note that in the right hand expression of eqn. 2 the radius is given in meters, speed in km/h and equilibrium cant in millimeters. This means that the constant $C \approx 11.8$ mm · m · h²/km². Cant deficiency (I) exists if the applied cant is lower than D_{EQ} and is a measure of the additional cant needed to obtain $D = D_{\text{EQ}}$. Cant deficiency is proportional to remaining track plane acceleration (eqn. 3)

$$I = D_{\text{EQ}} - D \quad (3)$$

If the cant deficiency is negative, i.e. $D > D_{\text{EQ}}$, the difference is called cant excess (E). In this case there will be an unbalanced lateral force in the running plane and the resultant force will move towards the inner rail of the curve. This can give problems for slow and heavy trains since a lower speed generates a higher cant excess. Therefore a limit for maximal cant value at speed v_f is needed (eqn. 4).

$$D \leq E + C \cdot \frac{v_f^2}{R} \quad (4)$$

Applied values in calculations are $E = 110$ mm and $v_f = 90$ km/h (Andersson and Persson, 2006). In practice, this limitation has influence in curve radii from 1500 m and up. Additionally there should also be a limitation on cant deficiency with regard to crosswind. In the speed profile calculations maximal cant deficiency is $I_{\text{lim}} = 300$ mm up to 225 km/h (v_{lim}). Over this speed allowed cant deficiency decreases with 1 mm/km/h (k) according to Andersson and Persson (2006). Depending on considered alternative, the maximal cant deficiency can be lower than I_{lim} . Fig. 18 shows the principles for setting cant deficiency value for a specific curve radius.

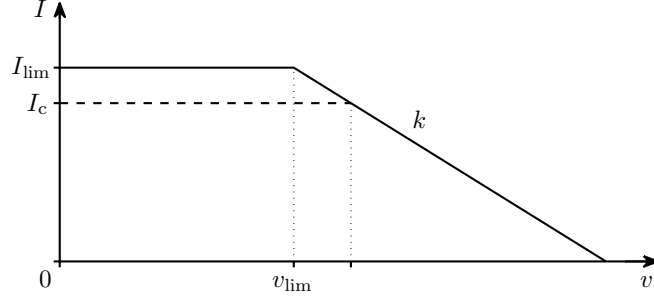


Figure 18: Allowed cant deficiency with regard to maximal case specific value, e.g. I_c , and to crosswind ($v > v_{lim}$).

Allowed cant deficiency is determined by solving for I_{cw} in eqn. 5 and taking the smallest value of I_{cw} and I_c . With this established the maximal curve speed is calculated with eqn. 6. Case specific cant value is used considering the limitation in eqn. 4.

$$I_{lim} + kv_{lim} - I_{cw} = \sqrt{\frac{k^2 R(D + I_{cw})}{C}} \quad (5)$$

$$v = \sqrt{\frac{R(D + \min\{I_c, I_{cw}\})}{C}} \quad (6)$$

Between a straight track section and circular curve or between two circular curves with different radius there is usually a need for transition curves. These are needed in order to achieve a gradual change of cant and curvature for comfort reasons. One characteristic is that the cant is changing as a function of the longitudinal position of the track and normally the mathematical form for a cant transition follows the mathematical form of the curvature which can be linear or non-linear. Transition curves have speed dependent limits for maximal rate of change of cant and cant deficiency (Kufver, 2000). In this study transition curves are omitted. In practice two cases can occur if a transition curve exceeds maximal rate of change for cant and cant deficiency. Either the speed must be reduced or the track geometry must be adjusted. The latter may require a major reconstruction.

6.3 Speed profiles

Curve speed calculations are not done for all existing curves. The aim is to find sections where a constant speed can be used. Lines with many curves or variation in curve radii usually give profiles with frequent speed changes. On lines with large curve radii or a high share of sections with straight track, a high constant speed is used on longer distances. The benefit of raising an already high top speed with for example 10% over a short distance is usually small.

This principle is used when new profiles are designed. Line sectioning is mainly based on existing speed profiles. Where applicable sections are split in two or more parts. The curve with the smallest radius is limiting the maximal speed and this is used in the calculations. Speeds are not changed close to larger urban areas and on larger stations where other limitations can apply. Also other circumstances can limit train speeds, e.g. geotechnical.

Running time calculations assume that main tracks are used where possible which means that speed limitations from using diverging tracks in turnouts are avoided. However, at some positions on the analyzed lines turnouts are passed in diverging positions and current speed restrictions are obeyed. Table 3 presents combinations of cant and cant deficiency used in the calculations (Andersson and Persson, 2006). The first two are existing profiles and considered as reference cases. Category B is for trains with no carbody tilt (P1) and category S for trains with carbody tilt (P2 and P3).

Table 3: Reference and computed speed profiles with cant and cant deficiency values (mm)

Profile	Cant (D)	Cant def. (I)	Information
P1	150	150	Reference category B trains
P2	150	245	Reference category S trains
P3	150	245	Carbody tilt
P4	160	165	No carbody tilt
P5	160	245	Carbody tilt
P6	160	275	Carbody tilt
P7	160	300	Carbody tilt

Speed profiles in Sweden have a base profile for category A trains. Almost all freight trains and some passenger trains belong to this category. Trains (vehicles) with track friendly running gear (radial steering bogies) are allowed to keep 10–15% overspeed relative to profile A on many lines. These are called category B trains. A maximal overspeed of 25–30% is allowed for track friendly trains with carbody tilt (category S). Currently only X2 trains have this functionality in Sweden.

A vehicle is considered track friendly if it causes low maintenance costs on the track and on the vehicle itself. In particular this gives advantages on existing non-perfect tracks. Speeds are rounded down which means that actual overspeed normally is less than 10/30%. There are also other technical issues related to the track which can limit the overspeeds. In addition, many sections both with low and high speeds, have the same top speed for all trains due to other reasons than train category.

Maximal curve speed for some of the fictive profiles is illustrated in fig. 19. Profiles P5–P7 differ in the range of 0–15 km/h, while P4 get clearly lower speeds caused by the significantly lower cant deficiency value according to table 3. There is no difference between profiles P5–P7 in the higher speed range due to the crosswind limitation. Speeds are also rounded down to nearest five multiplicative.

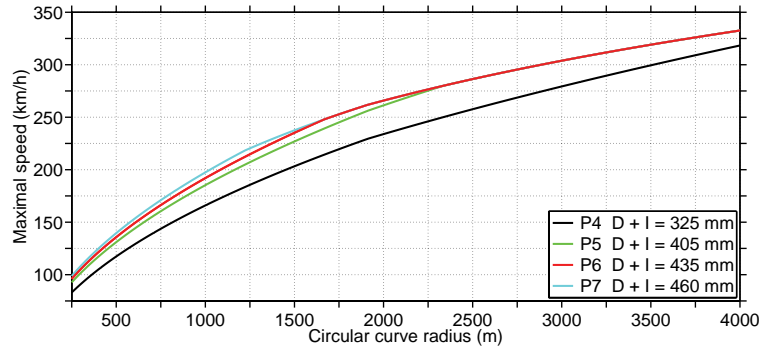


Figure 19: Maximal curve speed for different equilibrium cant values.

An example of where and to what extent a speed increase is possible is given in fig. 20. It describes the route Stockholm–Umeå and shows current speeds for B-trains and S-trains (P1 and P2) and the calculated speed profiles intended for trains without (P4) and with carbody tilt (P6). Since the speed difference between profiles P5–P7 is small, P6 is chosen as comparison profile. Speed profile P6 is also judged to be most realistic among other things with regard to travel sickness. Fig. 20 shows how the old winding section differs clearly from the converted or new sections, where speeds up to 300 km/h are possible with profile P6. Härnösand–Umeå has mostly radii around 2000–3000 m. It can be seen that the split between new/upgraded and older sections is around 50%.

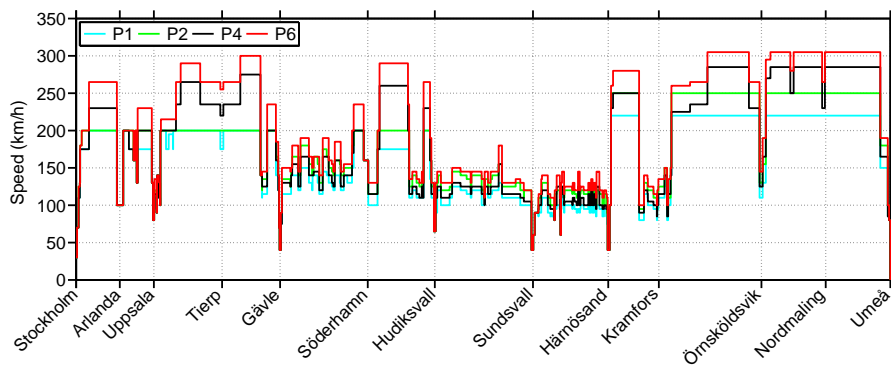


Figure 20: Current and calculated speed profiles for the route Stockholm–Umeå (East Coast, Ådalen and Bothnia Line), distance 713 km.

6.4 Train characteristics

Performed run time calculations are made for trains with varying characteristics regarding maximum permitted speed (top speed), start acceleration and power to weight ratio. The aim is to analyze the impact on run times regarding these characteristics and stopping patterns for a number of different lines. Levels used are specified in Fröidh (2006).

Common for all trains are weight 360 tons, length 155 m and an addition for rotating masses with 5.6% (dynamic mass supplement). Braking characteristics are analyzed as uniform deceleration of 0.6 m/s^2 , which is so-called comfort braking. Deceleration is not linear in reality, but can be reasonable well approximated with a constant value. It is assumed that at least half of the axles are powered, meaning that the adhesive mass is at least 180 tonnes. Adhesion is not considered to affect the short-duration tractive power under normal conditions.

Traction force diagrams are defined for the modeled trains. Desired starting acceleration (a_s) determines the level of starting tractive effort. Available tractive power (P) is limited above a certain speed, therefore the tractive force (F) is reduced. Traction force diagrams are defined by eqn. 7 and 8 where train mass including dynamic mass supplement is used (m_e).

$$F = m_e \cdot a_s \quad (7)$$

$$P = F \cdot v \quad (8)$$

Aerodynamic properties of a train are important and, if well designed, can reduce required traction force, energy consumption and run times. Aspects to be considered are for example air drag, impulse resistance and aerodynamic loads caused by crosswinds and the slipstream along the train.

Aerodynamically induced forces are typically proportional to speed or the square of speed. The sum of mechanical and aerodynamic resistance (D_{ma}) is approximated with eqn. 9 where $A = 2400 \text{ N}$, $B = 60 \text{ kg/s}$ and $C = 6.5 \text{ kg/m}$ are vehicle constants (Fröidh, 2005).

$$D_{\text{ma}} = A + Bv + Cv^2 \quad (9)$$

Gradients contribute to total resistance ($D = D_{\text{ma}} + D_g$) and will, depending on whether the train is facing a downhill ($D_g < 0$) or uphill ($D_g > 0$), reduce or improve acceleration characteristics. Available traction force is $F - D$ and the acceleration is calculated for a sufficiently small speed interval (eqn. 10).

$$a_i = \frac{(F_i + F_{i+1}) - (D_i + D_{i+1})}{2m_e} \quad (10)$$

Traction force and acceleration diagrams for reference trains and some variants of the Green Train appear in fig. 21. The power assumed relates to the maximum short-duration power used for a limited time during full acceleration or electric braking. Assumed that power output is constant, the start acceleration is affected by changing the gear ratio between motor and wheels. The existing power can be used either to provide high tractive effort when starting and a low top speed or low tractive effort when starting and a high top speed (Östlund, 2012). In this study top speed, power to mass ratio and start acceleration define the traction force characteristic of a train. This gives, supplemented by the running resistance, the acceleration performance on level track. The effect of gear ratio settings is not considered.

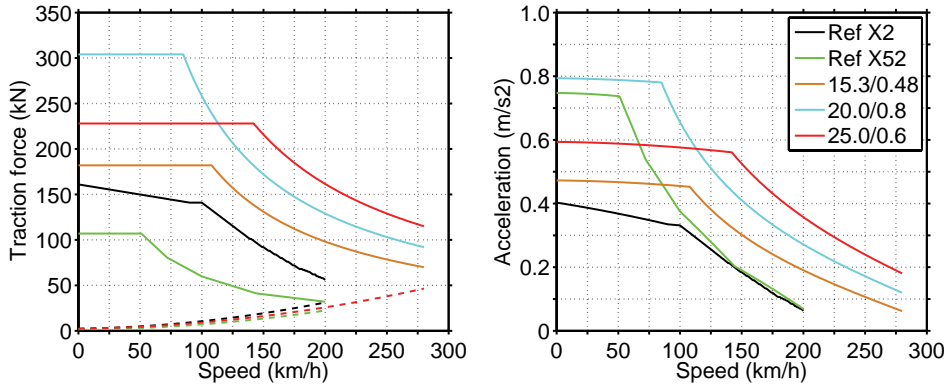


Figure 21: Traction force, total resistance with $D_g = 0$ (dashed) and acceleration diagrams for reference trains and combinations of power output (kW/ton) and starting acceleration (m/s²).

Power requirements are usually defined by desired residual acceleration, i.e. acceleration level at top speed. The difference between obtained residual acceleration is clearly seen in fig. 21. Gradients affect the acceleration and can, depending on the distribution of negative and positive gradients, increase or decrease acceleration times. Trains which have a low residual acceleration can face problems in reaching or keeping top speed even at relatively moderate positive gradients. Fig. 22 shows acceleration times for selected trains on a flat track and with gradient. In this case a power to ton ratio of 15.3 kW/ton almost doubles the acceleration time for reaching 280 km/h with a 6‰ gradient compared to the flat track case.

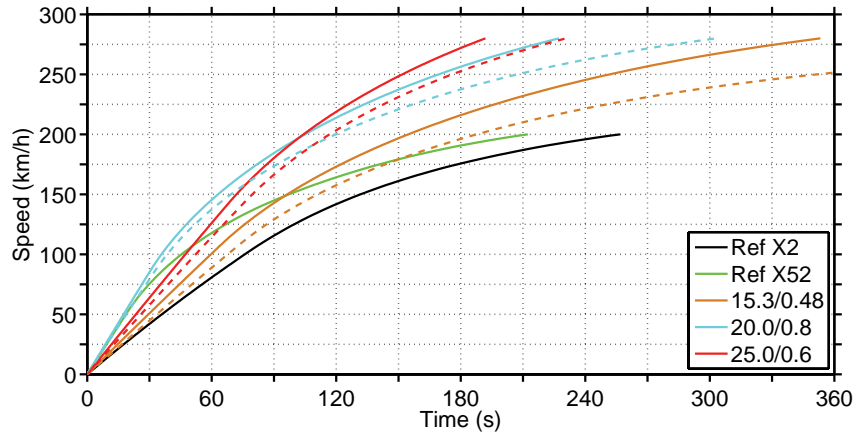


Figure 22: Acceleration time for reference trains and combinations of power output (kW/ton) and starting acceleration (m/s^2), flat track (solid) and gradient 6‰ (dashed).

6.5 Run time calculations

The simulation tool RailSys is used for run time calculations. For this purpose it is sufficient with a simplified infrastructure model, i.e. main tracks, stopping positions, signals and gradients. However, it is essential to set correct distances and speed profiles. Lines are modeled according to the situation year 2006–2007 including on-going construction projects of which some have been completed until year 2015. Vehicle models are defined according to the steps described in section 6.4 and assigned to speed profiles.

Cases with different stop patterns are defined for the studied relations (see Sipilä, 2008). Generally most of them have one stop pattern with few or no intermediate stops and one pattern including several stops. Station dwell times include passenger exchange times and in some cases additional time for train meetings on single-track sections. Initially the top speed for trains is varied in three levels, power to mass ratio in five levels and start acceleration in four levels according to table 4.

Examples of run times for some of the configurations are presented in table 5 showing total run times for reference trains and for trains with different maximal permitted speed (MPS), equilibrium cant (D_{EQ}) and power output. Dwell times are not included, i.e. braking and acceleration for station stops is considered but the standstill times are zero. The speed profiles for reference trains correspond to P1 and P2 in table 3). The variant trains follow speed profiles P4 and P6. Run times are rounded upwards to full minutes.

Table 4: Train configurations (the number of combinations is 60 but depending on the case some of these are not considered), top speed 320 km/h and $P = 30.0$ kW/ton used on cases which include Eastern Link and/or Götaland High-speed Line

Parameter	Levels				
Top speed (km/h)	250	280	320		
Power output/mass P (kW/ton)	10.8	15.3	20.0	25.0	30.0
Start acceleration a_s (m/s ²)	0.48	0.6	0.8	1.0	

The reason for different distances on equal origin and destination pairs is that different future routes are considered. For example, train running Stockholm–Gothenburg with 9 stops are routed via Nyköping. This is not the case for non-stop alternative. It is clear that an increase in top speed from 250 to 280 km/h give small time gains in relation to total time and distance in the examples. Profile P4 is aimed for non-tilting trains and in order to fully benefit from the higher top speed, curve radii must increase from around 2300 to 3100 m (fig. 19). This is mostly obtained on recently built lines. The situation is similar for trains with carbody tilting running on profile P6, an increase of top speed gives relatively small time gains in the presented cases. Changing from profile P4 to P6 (increasing allowed cant deficiency) gives the most significant improvements. The same effect is also seen for the reference trains comparing profile P1 and P2.

Table 5: Total run times for reference trains and variants of the Green train. Run times include 3% allowance as well as acceleration and braking for scheduled stops but exclude dwell times. Distance in kilometers, equilibrium cant (D_{EQ}) in millimeters, power output (P) in kW/ton and run times in hours and minutes. The Green train variants have start acceleration 0.8 m/s^2 and constant braking 0.6 m/s^2 . Table 5a represents, with some exceptions, existing lines. Table 5b represents relations that, to a varying degree, use future high-speed lines.

5a

Route	Distance	Stops	D_{EQ}	P	\rightarrow	Reference						MPS 250 km/h						MPS 280 km/h							
						300 P1		395 P2		325 P4		435 P6		15.3		25.0		15.3		25.0		15.3		25.0	
						X52	X2	X2	X2	2:42	2:41	2:24	2:23	2:41	2:40	2:23	2:22	2:41	2:40	2:23	2:22	2:41	2:40	2:23	2:22
Stockholm–Gothenburg	455	0				2:59	2:38	2:42	2:41	2:24	2:23	2:22	2:41	2:40	2:23	2:22	2:41	2:40	2:23	2:22					
Stockholm–Gothenburg	455	8				3:09	2:55	2:53	2:51	2:37	2:35	2:33	2:52	2:50	2:36	2:33	2:52	2:50	2:36	2:33					
Stockholm–Copenhagen	657	3				4:27	4:04	4:02	4:00	3:41	3:38	3:36	4:02	4:00	3:39	3:36	4:02	4:00	3:39	3:36					
Stockholm–Copenhagen	657	13				4:36	4:17	4:11	4:08	3:51	3:47	3:45	4:11	4:08	3:49	3:45	4:11	4:08	3:49	3:45					
Gothenburg–Copenhagen	342	3				2:15	2:15	2:05	2:03	2:02	2:01	2:00	2:04	2:02	2:00	1:58	2:04	2:02	2:00	1:58					
Gothenburg–Copenhagen	342	14				2:29	2:34	2:21	2:18	2:19	2:16	2:14	2:21	2:18	2:18	2:14	2:21	2:18	2:18	2:14					
Stockholm–Umeå	713	10				5:23	5:03	4:53	4:51	4:30	4:27	4:25	4:50	4:47	4:25	4:21	4:50	4:47	4:25	4:21					
Gävle–Umeå	530	15				4:22	4:08	4:00	3:42	3:57	3:38	3:35	3:59	3:39	3:35	3:35	3:59	3:39	3:35	3:35					

5b

Route	Distance	Stops	D_{EQ}	P	\rightarrow	Reference						MPS 250 km/h						MPS 320 km/h							
						300 P1		395 P2		325 P4		435 P6		20.0		30.0		20.0		30.0		20.0		30.0	
						X52	X2	X2	X2	2:11	2:09	2:10	2:08	2:11	2:09	2:10	2:08	2:11	2:09	2:10	2:08	2:11	2:09	2:10	2:08
Stockholm–Gothenburg	467	0				2:35	2:35	2:27	2:26	2:25	2:24	2:23	2:27	2:26	2:25	2:24	2:27	2:26	2:25	2:24					
Stockholm–Gothenburg	470	9				2:52	2:59	3:29	3:16	3:27	3:14	3:05	3:29	3:16	3:27	3:14	3:29	3:16	3:27	3:14					
Stockholm–Copenhagen	629	3				3:58	3:46	3:41	3:29	3:39	3:27	3:19	3:41	3:29	3:39	3:27	3:41	3:29	3:39	3:27					
Stockholm–Copenhagen	632	13				4:09	4:03	3:18	3:08	3:17	3:07	3:07	3:18	3:08	3:17	3:07	3:18	3:08	3:17	3:07					
Stockholm–Copenhagen	622	3				3:49	3:40	3:09	2:57	3:07	2:54	2:54	3:09	2:57	3:07	2:54	3:09	2:57	3:07	2:54					

6.6 Conclusions

The impact train characteristics have on run times varies depending on the speed profile. Lines with high continuous top speeds impose other requirements on trains than lines with shifting speeds. To make use of shorter sections with higher speeds a high power output is needed since accelerations in higher speed ranges are time consuming. Requirements on residual acceleration is a typical constraint on power output, especially for uphill gradients. This study assumes that an increasing starting acceleration, everything else being equal, does not limit top speed. Thus, acceleration time to full speed always decreases if start acceleration is increased. Considering starting acceleration and power output, usually the first step from 0.48 to 0.6 m/s² and 10.8 to 15.3 kW/ton gives the most observable time gains.

According to fig. 19 top speed level is only marginally improved by increasing cant deficiency from 275 to 300 mm. The lower limit is also a normal value for tilting express trains in Europe. Cant deficiency only affects comfort and the values are considered to give an acceptable centrifugal force in the trains' passenger compartment given carbody tilting. On almost all of the investigated lines changing from non-tilting to tilting speed profile gives the highest time gain.

On newly built and future lines this has less effect, instead top speed levels show a high elasticity. The biggest advantage with tilting trains is observed on lines with small or medium curve radii, this advantage decreases on lines with large radii. Stockholm–Gothenburg is a typical relation where trains without tilting functionality are not an option considering the run time differences between profile P4 and P6 or between the reference trains. Similar results are presented in Persson (2010).

As mentioned earlier, braking is not varied. However, increased deceleration can give similar improvements as increased start acceleration and is useful on services with many stops and on lines with shifting speed profiles. Acceleration and braking performance are also important in real operations since they can be used to reduce delays. These characteristics can in general be kept at a moderate level not to decrease the passenger comfort.

7 Multiple timetable and infrastructure analysis (Paper E and F)

A common approach when initializing a railway simulation study is to start from one timetable, an existing or assumed one, and make changes in order to get one or several evaluation scenarios that can be compared to the base line scenario. A study can simultaneously assess timetable and infrastructure changes or just concern one of them. Depending on how much is known about future conditions, it may be sufficient to design a few timetable scenarios for operational simulations, evaluate these and draw assumptions. On the other hand there is a risk of bias if the assumptions are uncertain.

Consider a situation where the aim is to evaluate effects of capacity improvements, e.g. an infrastructure expansion. This can be used in any combination of increased number of trains, decreased travel times and improved operational performance (reduced delays). An assessment of an infrastructure expansion will give different results depending on how the increased capacity is used. One timetable can e.g. be good from one viewpoint and less good from another viewpoint. Comparing a relatively poor timetable on one infrastructure and a good timetable on another infrastructure is misleading. This can in part be avoided by including multiple timetables in the analysis which can be evaluated and compared with respect to some performance measures.

An operational performance measure can be obtained by running operational simulations on respective timetable. This gives possibilities to compare static and dynamic timetable performance. The static performance can e.g. be scheduled delay, i.e. the difference between nominal and obtained run time, or relate to heterogeneity describing the temporal usage of line sections and stations etc. The dynamic performance reflects results from operational simulations, i.e. simulations with stochastic delays.

Fig. 23 illustrates the concept. Multiple timetables are generated and ranked according to a static performance measure (B_1). A subset of these are evaluated with respect to a dynamic performance measure. The dots represent specific timetables in the subset (A_1). Assume that another setup is used, e.g. a different infrastructure or frequency of trains, then a new subset (A_0) of timetables can be evaluated in the same way and compared to the first scenario. The subset areas may naturally overlap.

Assume that the aim of a study is to compare two different infrastructure expansions and relate them to an existing one, i.e. the do-nothing case. There is potentially a great risk that the outcome of a cost-benefit analysis based on only one timetable in each case gives little insight. If the timetables in fig. 23 are designed so that one scenario is low whereas the other scenario is high on both static and dynamic performance with respect to the subsets of timetables, the difference in performance will be misleading.

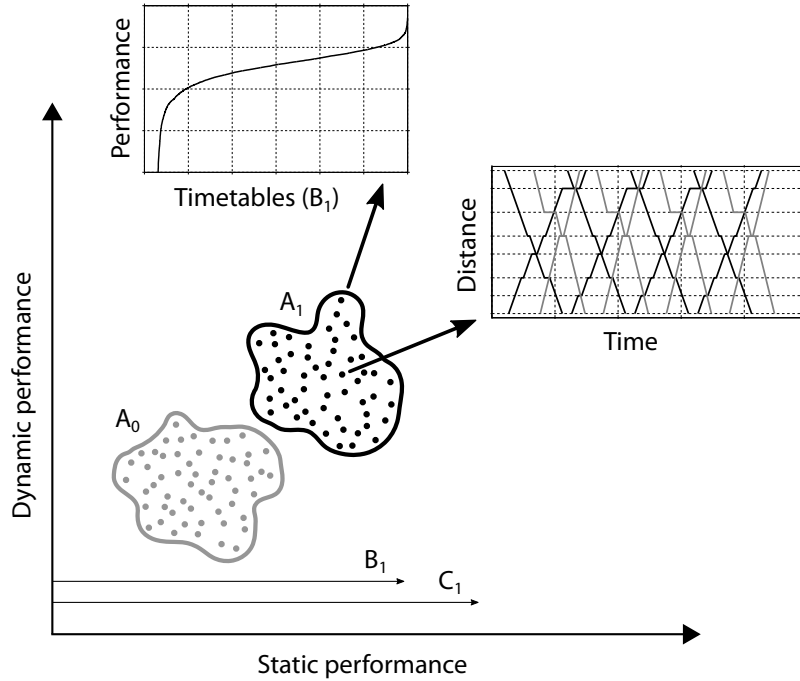


Figure 23: Concept for analyzing performance measures for multiple timetables on different infrastructure and/or train frequency scenarios.

The subset timetables A_1 can for example represent the n best timetables in another set B_1 , i.e. $A_1 \subseteq B_1$. The term *best* relates to the ranking with respect to a chosen measure or mix of measures on the static performance axis. The rest of the timetables in set B_1 , not included in A_1 , will by definition be positioned on the left side of subset A_1 if a higher axis value is considered better than a lower. It is important to emphasize that set B_1 is a subset to the set of all possible timetables (C_1) with respect to initiation parameters in each case, i.e. $A_1 \subseteq B_1 \subseteq C_1$. The timetables in C_1 are spread on the static performance axis, both to the left and right as well as overlapping the subset A_1 . Depending on what the static and dynamic measures consist of, there may be timetables that have a poor static performance but a high dynamic performance and vice versa.

The idea is that set C_1 can be represented by a sampled subset B_1 . The reason for using sampling is to get a manageable problem size for simulation. Set C_1 is formulated as departure time combinations for respective train groups and initiation locations. When it comes to comparing two scenarios (e.g. A_1 and A_0), one measure could be to calculate their respective mean values on respective axis. This could give a good perception if the dots are relatively concentrated. However, a dispersion measure should be added to better capture the effect of the subset shapes and sizes.

7.1 Nominal timetables

Constructing timetables for a line or network manually implies that all requests for train paths should be considered, trains (individual or patterns) laid out and conflicts handled appropriately. The blocking time principles and other possible restrictions should be included in this process, meaning that a truly conflict-free timetable is obtained. In addition to this, sufficient time separation between train movements need to be considered.

The nominal timetable represents the step where trains are laid out in the system but no conflict management is performed. One way of interpreting the nominal timetable is to say that all trains have their requested start times and paths. This is of course a fictive scenario since a nominal timetable will rarely equal an operational timetable, except in cases with sparse traffic or excessive infrastructure resources. Depending on the amount and size of conflicts, the final operational timetable may look quite different from the nominal one. Solving one conflict will often lead to new conflicts elsewhere in the system and so on.

Creating multiple nominal timetables is achieved by varying all requested train departures combinatorially and subsequently introduce them in the timetable. Each combination will thus give one unique nominal timetable, the difference from one combination to another may be small. The nominal timetables can then be conflict-managed by simulating them and result in operational timetables with varying quality. These can then be ranked by introducing one or several performance measures. Typically they would in some way reflect to the deviations from the nominal timetables but also other factors may contribute. Scripts are used for creating departure matrices for different locations. Input parameters that need to be considered are:

- Train group identifiers
- Departure frequency per group
- Minimum headways (time) between groups
- Time resolution
- Cycle repetition
- Sampling

Train group identifiers are needed to keep track of different trains in the database and during the assignment process. Departure frequency specifies how often trains should depart in respective group. Minimum headways are set in a matrix and used for excluding unrealistic or other undesirable solutions. The time resolution parameter defines the time slot length, typically one minute. A lower resolution, i.e. a higher value, reduces the number of possible combinations.

A total cycle time is calculated before the combination matrices are generated. The cycle time is indirectly given by the combination of departure frequencies which should be chosen with some care, i.e. avoiding unrealistically high cycle times. Cycle repetition is an integer specifying how many times a full departure cycle should be repeated, i.e. how long time the nominal timetables will cover. If sampling is active, the specified number of combinations are sampled uniformly from the total number of combinations. Both sampling and time resolution can be used to get a manageable number of combinations.

The symbolic matrices below show the principles for creating combinations at different train initiation locations $\{1, 2, \dots, p\}$ and with train groups $\{1, 2, 3, \dots, n\}$ distributed on the locations. The time resolution parameter together with the cycle time define the number of columns in each matrix. This is identical for all locations. The number of combinations (rows) in each location, depends on the number of train groups, the number of possible time slots (columns) and also on the defined minimum headways. The total number of combinations for a setup is given by multiplying the number of rows in the matrices.

$$\begin{array}{ccc}
 & 1 & 2 & & p \\
 \begin{bmatrix} 1 & 2 & 0 & \dots & 0 & 0 \\ 1 & 0 & 2 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & 1 \end{bmatrix} & \begin{bmatrix} 3 & 0 & 0 & \dots & 0 & 0 \\ 0 & 3 & 0 & \dots & 0 & 0 \\ 0 & 0 & 3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 3 \end{bmatrix} & \dots & \begin{bmatrix} n & 0 & 0 & \dots & 0 & 0 \\ 0 & n & 0 & \dots & 0 & 0 \\ 0 & 0 & n & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & n \end{bmatrix}
 \end{array}$$

For a single-track line with equal traffic flows in both directions, the number of combinations for a setup with three train groups and corresponding frequencies $\{60, 120, 180\}$ gives 3422 unique combinations for the first direction. A reduction is possible for the first location since departures for the first train group can be fixed, varying this would give otherwise equal combinations except that they are time shifted. The second location or opposite direction gives around $1.232 \cdot 10^6$ combinations. Merging the directions means that the first is multiplied with the second, which in this case gives a very large number.

If we assume that the first two groups are passenger service and the last one is freight, some headway restrictions can be introduced. To avoid bunching and get a more attractive spread of departures from the initiation stations, a minimum headway of 20 minutes would be reasonable. Additionally the freight train is given at least 5 minutes headway to a preceding passenger train and 10 minutes in the reverse situation. The number of combinations is now 672 and 241920 for the two directions. Although a significant reduction compared to the first case, the total number of combinations is still large.

The difference between successive combinations can be small since one train group may have been shifted with only one minute. It is therefore reasonable to expect that the simulation of nominal timetables, in these cases, can give almost similar solutions. However, at some point a step effect occurs resulting in a changed train meet or overtaking location. Using a lower time resolution (longer time slots) reduces the probability of getting many similar solutions.

Sampling can be applied to get a manageable number of departure combinations for simulations. Another possibility is to generate combinations for two train groups, run simulations and rank the solutions. After that, a number of departure combinations can be fixed with respect to the trains already simulated and more train groups can be combinatorially inserted. Train groups in the first simulation run should have a higher dispatching priority than the subsequently inserted groups.

The total simulation time will depend on the number of combination cycles, the size and characteristics of the line/network and the number of trains which relates to the dispatching effort. After the necessary preparations are done in the simulation software and scripts are run for providing requested number of departure combinations for a specified setup, a database is created from which the actual departure assignment file (XML-format) is generated.

The principle with a nominal timetable is illustrated in fig. 24. It shows a single-track line where meets and overtakings can be executed at locations with sidings (horizontal lines in figure). Different train groups are defined by run times including scheduled stops. The nominal timetable, described by one departure combination, includes conflicts. The objective is to run simulations on multiple nominal timetables and get operational timetables that can be evaluated with respect to performance measures. Running a nominal timetable is similar to an unscheduled operation, hence the trains must be conflict-managed as they traverse the network.

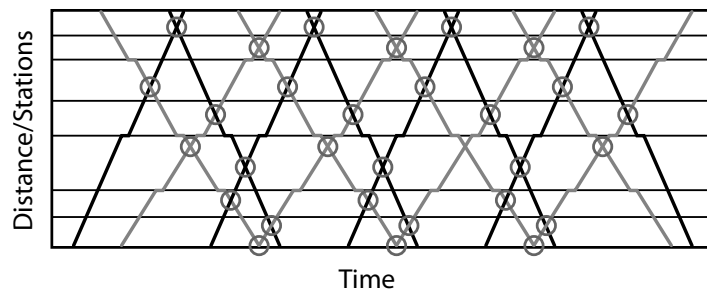


Figure 24: Nominal timetable on single-track line with conflicts (circles).

The input parameter *cycle repetition* is important since it, in conjunction with the cycle time, decides the total time for a nominal timetable. A sufficient time

span of fully developed traffic is necessary for the evaluation, i.e. the warm-up and cool-down periods should be observed. The concept for creating nominal timetables follows the same principles regardless of if a pure single-track line, a double-track or multi-track line or a network is studied. The main difference is that the number of unique departure combinations increases with the number of initiation locations. Sampling will most likely be necessary in many of the cases that are interesting to study, real or fictive ones.

7.2 Operational timetables

Following the simulation, data is exported, postprocessed and fed to the database. Required information is cycle numbers, timestamps and realized routing through stations. The evaluation relates to the nominal timetables and will give a measure as to what degree the requested train paths have been fulfilled. There will typically be a distribution of timetables in which a small share is interesting for further investigation and operational simulation. How large this share will be in relation to the total number of timetables depends strongly on the traffic and infrastructure setup.

A typical and relatively simple measure of performance is the scheduled delay, i.e. the difference between the nominal and realized run time for each train. On a single-track line, most of the scheduled delay derives from waiting time at stations where meets and overtakings are processed. The rest of the schedule delay is naturally caused by exceeded run times between stations due to train interactions, e.g. the lack of parallel entrance to stations meaning that one train has to wait until another train has entered the station taking safety time into account. On a double-track line with mainly unidirectional operation on each track, the interaction effects on the line may attribute to a larger share.

Scheduled delay can for example be given in seconds or minutes and normalized with a distance if preferred. It can also be a percentage of added time with respect to the nominal time. Fig. 25 shows an example of how the operational timetables are distributed. This particular case is a single-track line with 11 stations, including boundary stations, and a total length of 107 km. Two passenger train systems, each one in 60 minute intervals, run with difference in top speeds and scheduled stops. There is no run time allowance included in the nominal timetables for this case, run times are 39 and 49 minutes including stops. Time resolution is 2 minutes and minimum initiation headway 10 minutes.

The total number of combinations for the full set, given the input parameters, is 13230. Using a simple of 500 combinations gives the same curve shape. Based on this representation, a set of operational timetables could be chosen for further investigation. For example, if an average of 10% added run time is accepted there are around 140 timetables in the sampled set. Poor timing in meets, i.e. adverse departure combinations, and software limitations when it comes to handling multi train conflicts on single-track lines can cause high scheduled delays.

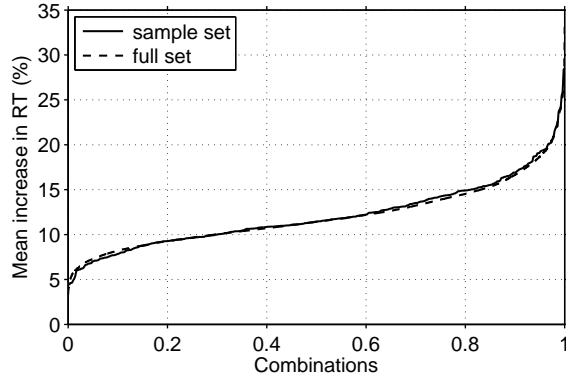


Figure 25: Distributions of simulated nominal timetables showing aggregated mean values for train groups, full set (13230) and sampled set (500) of combinations.

In fig. 25 the curves are aggregated on both train groups. If requirements on maximum deviation from the nominal run time are applied so that both groups must satisfy the condition simultaneously, the number of approved timetables may drop significantly. Fig. 26 illustrates how the mean percentage in added run time is distributed if the second train group is sorted with respect to the first one. If both groups must meet an acceptance level of 10%, the number of timetables decreases compared to fig. 25. These two groups can further be split into two subgroups each to be consistent with the initial combinatorial setup where two groups in the first direction are combined with two groups in the second direction.

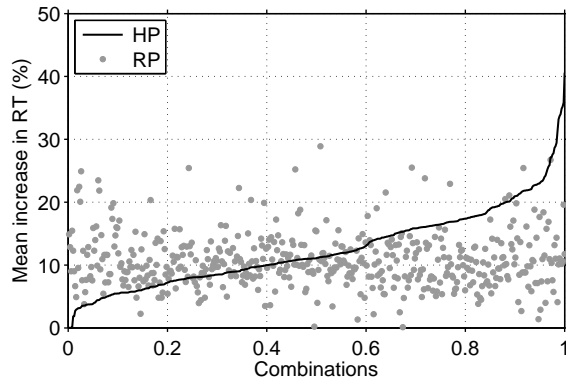


Figure 26: Distribution of simulated nominal timetables for two train groups, where the first one is sorted from smallest to highest and the second one follows the same sorting.

Another aspect for passenger train services is regularity, i.e. trains in a specific group have a certain regularity in departure and arrival times. This does not necessarily mean that the temporal distance between trains is equal but there should be a cyclic regularity. Applying synchronous simulation on nominal timetables and mixing train groups with different frequencies can potentially give irregular timetables for some train groups. The regularity property can be introduced in an evaluation of operational timetables. One way of assessing this is to compare the requested departure interval times with arrival interval times within a train group.

Fig. 27 depicts the distribution of operational timetables if a limit of 60 seconds is introduced on two passenger train groups, a mean value of the standard deviation for four groups with regard to the direction. Some timetables are perfectly regular, i.e. the value is zero. A moving mean consisting of 1% of the values gives an indication of the trend if sorted with respect to scheduled delay. The values have a high spread, regular timetables exist for low as well as high scheduled delay values.

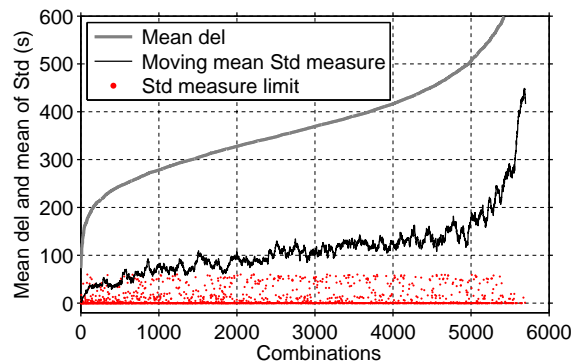


Figure 27: Mean values for standard deviations of the differences in nominal and realized arrival times, sorted with respect to scheduled delay. Distribution of values less than or equal than 60 seconds and a moving mean line for all timetables.

In reality there may be limit values on scheduled delay for several train groups in a timetable. Timetables which meet the requirements will then consist of the ones where all groups simultaneously have values less than or equal to their respective limits. The groups can be weighted differently, e.g. in order to give more priority to some groups. Fig 28 shows distributions of several parameters on a single-track line, same infrastructure configuration as before, with both passenger and freight trains introduced in sequence. In this case it means that the passenger train groups are simulated first, combinations giving the 30 best solutions are identified after which freight train departures are added combinatorially.

Sorting regional (RP) and freight train (FR) values according to the high-speed train curve (HP) and using moving mean provides an indication of possible correlation in increased run time. It is clear that the variation is relatively high, although an increasing trend is observed for train group RP. Freight trains show no correlation with respect to high-speed trains using this approach. Considering both passenger train groups as one group and sorting freight train values in accordance give similar results, i.e. no correlation trend is observable.

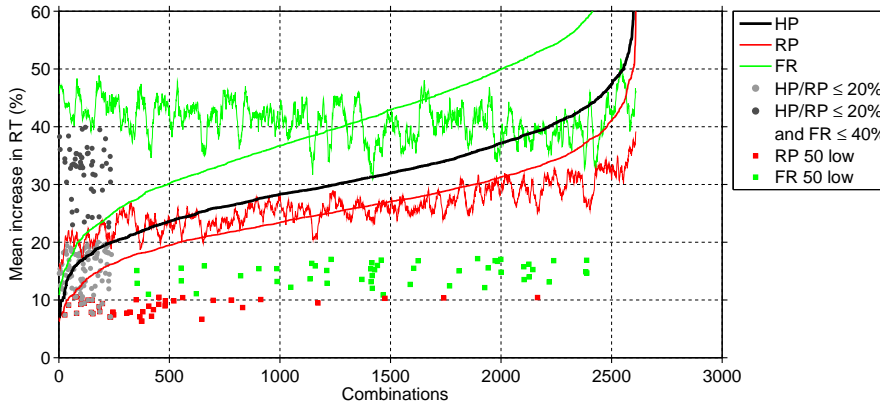


Figure 28: Mean percentage of increased run time, sorted individually for each group and according to high speed trains (HP) as a moving mean with sample window 1%. Timetables not exceeding specified limits (%) are indicated together with timetables giving the 50 lowest group values.

Assuming that some of these sequences are used as input to stochastic simulations in order to evaluate the effect of delay propagation, threshold values can be used in finding preferred timetable solutions. Figure 28 shows an example in which solutions satisfying requirements on maximum mean increase in run time are indicated. As a comparison, the figure includes the 50 lowest values for regional and freight trains. It is clear that simultaneously requiring low values for all three groups will hardly give any solutions under these conditions. Nearly all of the solutions giving the 50 lowest increased run times for freight trains lie in an area outside of $HP \leq 20\%$.

Giving some numbers, from a total of 2610 feasible timetables 240 satisfy the condition $HP \leq 20\%$ meaning that the percentage of increased run time is at most 20%. Adding the same condition for regional trains gives 122 timetables, i.e. $HP \leq 20\%$ and $RP \leq 20\%$. Approximately half of the 50 solutions giving the lowest increased run time for regional trains seem to lie in the region $HP \leq 20\%$. If these timetables are further filtered by also requiring $FR \leq 40\%$, 56 timetables remain. This exemplifies that even if the number of acceptable timetables considering only one group can be high, simultaneous requirements for several groups can give significantly lower number of timetables.

7.3 Stochastic simulations of operational timetables

A set of operational timetables can be simulated with stochastic perturbations (operational simulation), thus providing a view of their dynamic performance. This can in turn be related to one or several static performance parameters. A common measure of performance in operational simulations is on-time performance and average delay. Comparing initiation and final values indicate if and to what extent trains are able to recover from their initial delays. Imposing limit values, e.g. regarding on-time performance on final station, could further reduce the number of accepted timetables. Applied stochastic delays need to be realistic to provide credible simulation results. Using several sets of delays, e.g. a high and low split, provide information on the sensitivity to changes. Timetables that have good recovery ability would give a smaller difference in final delays compared to timetables with poor recovery ability.

A property that will have effect on the dynamic performance is the included allowance (slack) in run times. This would be set as a percentage of minimum run time in the nominal timetables discussed in section 7.1 and transferred to the operational timetables. Some or all of this allowance can then be made accessible in the operational simulations and thereby contribute to reduce delays. Fig. 29 gives an example of this effect and shows both the static and dynamic performance for 100 timetables with two passenger train systems reflecting the same infrastructure as used in fig. 25–28. An exponential distribution with a mean value of 3 minutes and truncated on 10 minutes affects 50% of the trains.

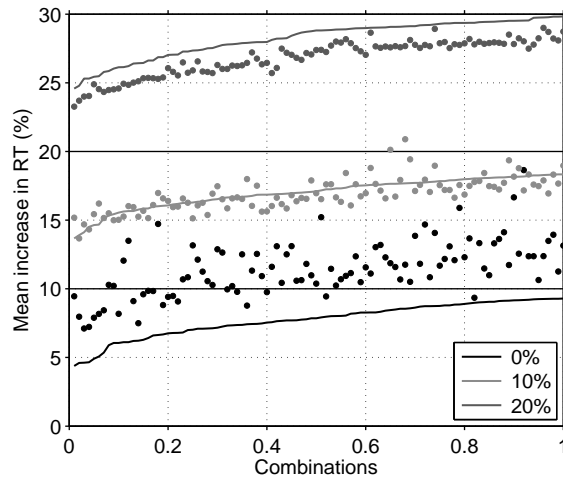


Figure 29: Simulated operational timetables with stochastic delays, shifted in y-direction with corresponding added allowance in legend.

The timetables are sorted according to the mean percentage increase in run time with respect to the zero percent case. The dots represent dynamic performance when stochastic initial and dwell delays are introduced. The interpretation is that if a dot lies below its corresponding static performance line the run times are on average shorter than the scheduled ones, i.e. there is ability to partly recover from delays. The opposite is true for dots placed above the line. As would be expected, the ability to recover improves with increased percentage of run time allowance.

Applying no allowance results in sensitive timetables and also a higher variation between timetables. This implies that some of the timetables are more preferable than others from an operational perspective. Since mean values are for four train groups (two in each direction) are used, it does not give the full picture and the variation can partly be related to this. This type of analysis where scheduled run time allowance is varied is useful if the objective is to find an acceptable percentage of allowance related to expected delays.

7.4 Assigning delays

In connection with selecting and configuring operational timetables a delay assignment routine is used. Definitions are read from a spreadsheet and time values (delays) are created and assigned to trains in the database along with position information. This setup works by pointing different types of delay distributions on respective train groups and defining percentages for trains affected, maximum values as well as direction and location information. It is also possible to use random locations. Theoretical distributions available in Matlab can be used with associated parameters. By controlling a random seed parameter, the same values can be replicated several times if no changes are done in the setup.

The routine can be expanded to use empirical distributions and to use a mix of stochastic and systematic delays, e.g. by assigning delays to consecutive trains on a location during a time interval or influencing certain trains on consecutive locations. The aim would be to model both infrastructure failures active for a certain time and vehicle related problems that may impair a train during the entire run. If available, historical data for these types of events can be used, thus enabling a more realistic modeling of delays. Assigning delays in this way allows for a more flexible setup than offered via the RailSys interface. Since the values are stored in the database, i.e. information on train ID, type of delay, value and simulation cycle, a more detailed analysis would be possible following a simulation. This could for example involve studying event chains for primary and secondary delays. The operational simulations require both timetable and delay files in XML-format, both are generated from the database.

7.5 Infrastructure as a variable

If the infrastructure is included as a variable in this type of analysis, it is possible to see what effect an improvement has on the timetable outcome. Changes in infrastructure is usually thought of in terms of improvements but it can also involve reductions in infrastructure, e.g. removing switches to reduce maintenance costs and the likelihood of failures. Adding a siding on a single-track line will often give operational benefits in disturbed conditions, even though it may be unused in the planned timetable.

A typical objective of using different infrastructure configurations is to e.g. investigate which improvements are needed for meeting certain requirements on traffic volume, scheduled and operational delays, etc. Assuming that investment costs can be determined, results from this type of study could then be used as input to a cost-benefit analysis (see e.g. Eliasson and Börjesson, 2014). The infrastructure changes considered can range from adding switches on stations to allow for more flexibility and simultaneous routing, to adding new tracks between stations, hence improving line capacity.

Fig. 30 shows an example where two traffic load levels are applied on infrastructure variants with equal total length (240 km). The difference moving from the first variant to the second is that sidings are spaced 10 km apart instead of 15 km on an otherwise single-track line. The following two variants have 25 and 50% of the total length configured as double-track. The stop patterns for passenger trains are equal as well as the departure combinations used in each case. Freight trains are operated with half or less than half the frequency of passenger trains.

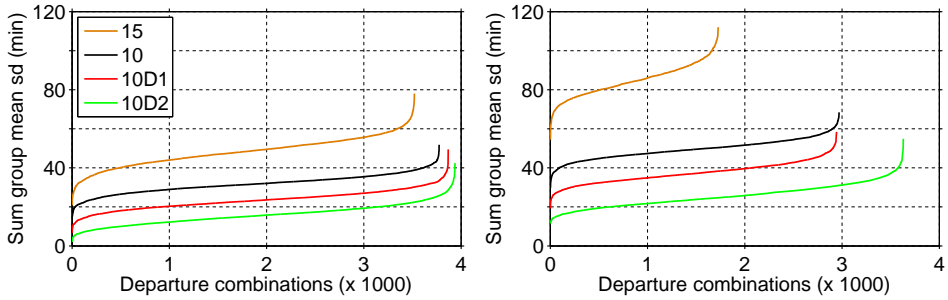


Figure 30: Sorted values for the sum of mean scheduled delay for passenger train groups on respective infrastructure alternative. Left and right figures have different departure frequencies.

The figure gives a view of the potential reductions in scheduled delays when the infrastructure is expanded. The second case with a higher number of departures (fig. 30 right) gives greater time savings comparing between variants than the first case. The scheduled delays are lower in the first case due to less number of trains, i.e. less number of conflicts originating from the nominal timetables.

By viewing fig. 30, one can observe that the number of operational timetables are different. Although the same number of nominal timetables are simulated, some of them are lost due to deadlocks. This becomes clear in the case with more trains. This is also the reason why the first infrastructure variant is not simulated with stochastic delays. The causes of deadlocks and their implications are discussed in section 7.7.

A small subset of the operational timetables can be used in operational simulations and evaluated with respect to operational delays. Choosing candidate timetables can, depending on which measures are considered important, be done in several ways. In this example the scheduled delays are used as a measure both considering passenger and freight trains, however the passenger trains are given more weight. Fig. 31 shows the outcome of summed operational exit delay per timetable for passenger trains.

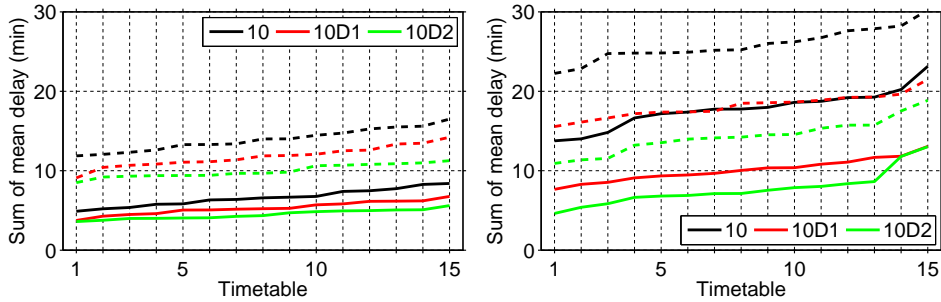


Figure 31: Sorted values for the sum of mean operational delays for passenger train groups. Left and right figures have different departure frequencies. Dashed lines represent higher initial delays compared to solid lines.

For the case with less trains (fig. 31 left) the spread in outcome is significantly smaller than in the other case. Moving from infrastructure variant with 25 to 50% share of double-track has a marginal effect on the outcome. The benefits of having a larger share of double-track is more clear in the case with a higher number of trains. This is an example of how increased line capacity is not only used for operating more trains in a timetable but is also important in operations. Since the sum of delays is used, the distribution on respective train group is not known. Furthermore, if the operational delays are compared with scheduled delays it is not necessarily the timetable giving the lowest operational delays that is ranked highest.

7.6 Infrastructure modeling

As mentioned earlier, the infrastructure has a strong influence when it comes to timetable construction. Making improvements on an existing infrastructure may remove some of the restrictions that e.g. prevents an increase of trains. In a railway infrastructure investment appraisal several alternatives may be of interest. This can e.g. involve locations of double-track deployment, new stations or entirely new lines. Sections 7.1 and 7.2 deal with a method to find a set of operational timetables which can be evaluated by performing stochastic simulations. The analysis can be further expanded by including the infrastructure as a variable. A study could for example be targeted to answer what infrastructure improvements are necessary to meet an on-time performance requirement given a certain traffic frequency.

The infrastructure in RailSys is defined by a node-link structure. Nodes have object properties, e.g. a signal, switch/turnout or stop location, links connect nodes and have attributes such as length, speeds, gradients and so on (see e.g. Radtke, 2008). This node-link structure can be designed by using the RailSys interface. The time required depends on the size and complexity of the model as well as user experience. However, there is no straightforward way of producing multiple infrastructure variants which may share some characteristics and differ on others.

Consider for example a case where the total line length is fixed but the aim is to vary the number, positioning and layout of stations as well as location of single, double or multi-track sections. Being able to design multiple infrastructure variants in a rapid and relatively simple manner will facilitate the inclusion of the infrastructure as a variable in simulation studies. The node-link structure used in RailSys is described by an XML file format. Consequently, the infrastructure can be generated from another application.

In order to speed up the process of generating infrastructure variants a model that uses spreadsheet definitions and scripts is created. The spreadsheets enable a relatively quick way of defining different station layouts with objects and distances, i.e. obtaining a library from which stations can be imported. Another spreadsheet is used for linking stations together and assign positions, thereby creating the line sections between them including signals and other associated objects.

Using scripts, the node-link structure can then be created for relevant object types, signal sections, station routes etc. The result is a ready-to-go infrastructure file which can be read by RailSys. In order to avoid loss of generality recursion is applied for path search, thereby not limiting the size and complexity of station layouts. Fig. 32 depicts a station layout spreadsheet formulation, in this case the node-link structure for a double-track station with one siding. This example includes some basic objects, there is however support for including several other object types that are available in RailSys.

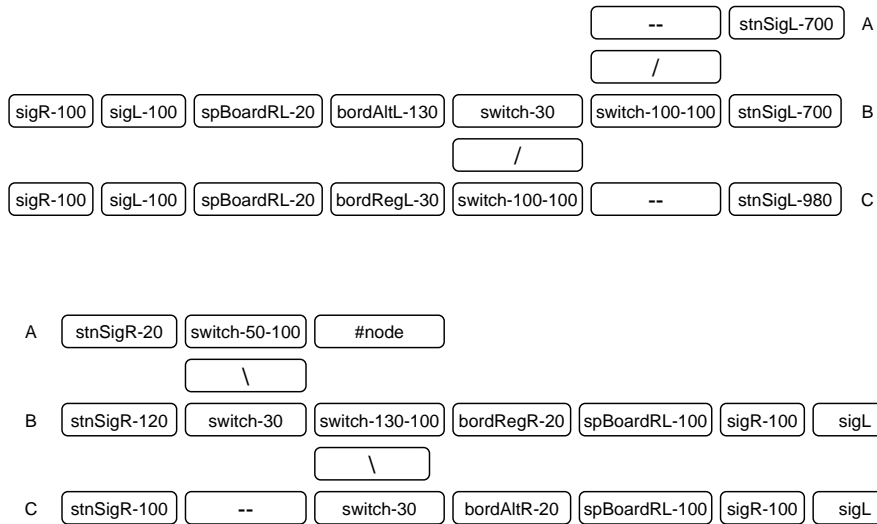


Figure 32: Example of a double track station spreadsheet definition.

Stations are given identifiers which are used in the line definition spreadsheet where line sections and stations are linked together. After the node-link structure is created for the whole infrastructure setup, signal and station routes as well as other necessary node and link attributes are set, e.g. to ensure that the correct dispatching method in RailSys will be used for single, double or multi-track lines. There is also a possibility to define if a station should allow simultaneous train entry or not. The positioning of stations implicitly defines the distances of line sections. The requested number of block signals are placed equidistantly. After all infrastructure data is compiled it can be described in the XML-format recognized by RailSys. The station definition in fig. 32 will be similar to fig. 33 when the XML-file is processed in RailSys.

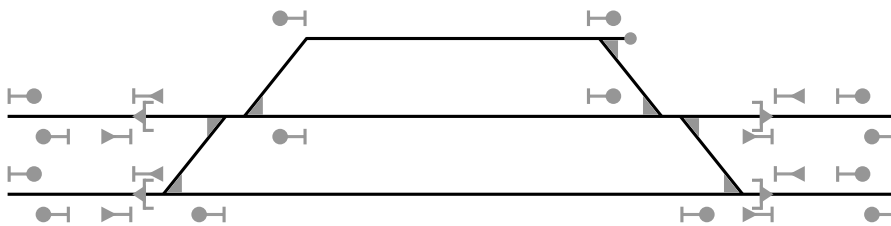


Figure 33: Node-link structure from fig. 32 similar to the layout in RailSys Infrastructure Manager module.

Defining or importing varying speed or gradients profiles is not possible at this stage. A flat speed for main tracks can be set and switch (turnout) speeds defined for each station in the line configuration input data. However, the infrastructure database is created in such a way that this can be implemented later. Furthermore, not all object types available in RailSys Infrastructure Manager module are supported. The focus has been to implement objects needed to model typical conditions in Sweden.

7.7 Handling of deadlocks

Dense traffic combined with stochastic delays, e.g. on complex infrastructure and for bidirectional operation, is to a varying degree prone to deadlocks in synchronous railway simulation (see Pachl, 2007, 2011). Even though the total traffic density can be relatively moderate, congestion may occur in an area covering a few stations and result in a deadlock. If no or only small delays occur, the scheduled train sequence may remain intact and no deadlocks are produced. However, if delays become large enough, the scheduled train sequence will most likely be changed which increases the probability of deadlocks.

Simulation of the nominal timetables tend to produce a varying number of deadlocks, especially in the presented cases due to the bidirectional operation (single-track). This is basically an unscheduled operation which resembles a scheduled operation with large and frequent delays. Although there are parameters for intervention time ranges and how much weight is given to train priorities in conflict management, the basic problem of deadlocks remains.

As just mentioned, the train runs can be viewed as completely unscheduled when the nominal timetables are simulated. Remembering that the purpose is to get operational timetables, it is up to the dispatching functionality in RailSys to conflict-manage trains regarding meets and overtakings. Some of the combinations (simulation cycles) result in deadlocks of the types described in fig. 5. The amount of deadlocks will normally increase when the number of trains operated per time unit increases. If the share of deadlocked cycles becomes too large it indicates that the number of timetable alternatives may not be particularly high for the simulated traffic load. Using another random seed for the sampling and simulating again should give roughly similar results regarding the number of deadlocked cycles in this case.

The operational simulations are setup by configuring several timetables in temporal sequences. Separation time with no activity is used in between timetables in order to ensure that there is no interference from one timetable to another. In this way multiple timetables can be evaluated in the same simulation instead of running one simulation setup per timetable. If the probability of deadlocks is sufficiently low, then all timetables can be considered active in each simulation cycle (replication).

If the probability of deadlocks is high, or in other words the probability for a timetable to pass a simulation cycle without a deadlock is low, then the overall probability of having a successful cycle for all timetables simultaneously will be even lower. One way of ensuring a sufficient number of evaluable cycles is to increase the number of requested cycles prior to a simulation. Another way is to deactivate all but one timetable in each cycle. This is done by assigning a high initial delay to all trains in these timetables so that they will not operate within the simulation time space.

There is no straightforward method in RailSys for train cancellation in relation to cycle numbers in a simulation. When delays are assigned to trains in the database during preparations for operational simulations, indicators for active and non active cycles for the different timetables are used so that the corresponding XML-file prepared for RailSys reflects this. When simulation results are processed into the database, the timetables are matched to their respective active cycles.

7.8 Conclusions

The objective of the presented method is to improve the reliability of studies of future timetables and facilitate the management of multiple infrastructure variants. This is possible by taking advantage of the simulation functionality in RailSys and using the developed routines for managing data prior to and after simulations. Using an asynchronous method for timetable generation (see e.g. Lindfeldt, 2011a) will most likely be faster in the actual simulation phase compared to synchronous simulation. However, the latter can provide more realistic timetables since it does not schedule full train runs in a strict order of priority. This method does not aim to find the optimal timetable. The purpose is to find a representative set of timetables that can be used in further studies and give an estimation of possible outcomes, both in a static and dynamic dimension.

Examples of single-track line applications are used here, but this method can be applied on a double-track line or a network with more than one line. The principles are equal although the number of initiation locations may increase, hence the number of unique nominal timetables increases as well. Since the problem size grows, the number of sampled combinations will relatively represent a smaller set of all nominal timetables. The uncertainty regarding this can partly be counteracted by drawing two random samples, simulate both sets and in the evaluation make an assessment whether the results are consistent or not. Fig. 34 presents the process step by step. Simulation of nominal timetables can also be performed in sequence by including e.g. passenger trains in the original setup and, based on the evaluation, select some of the nominal timetables and combinatorially add freight trains while locking the trains already included.

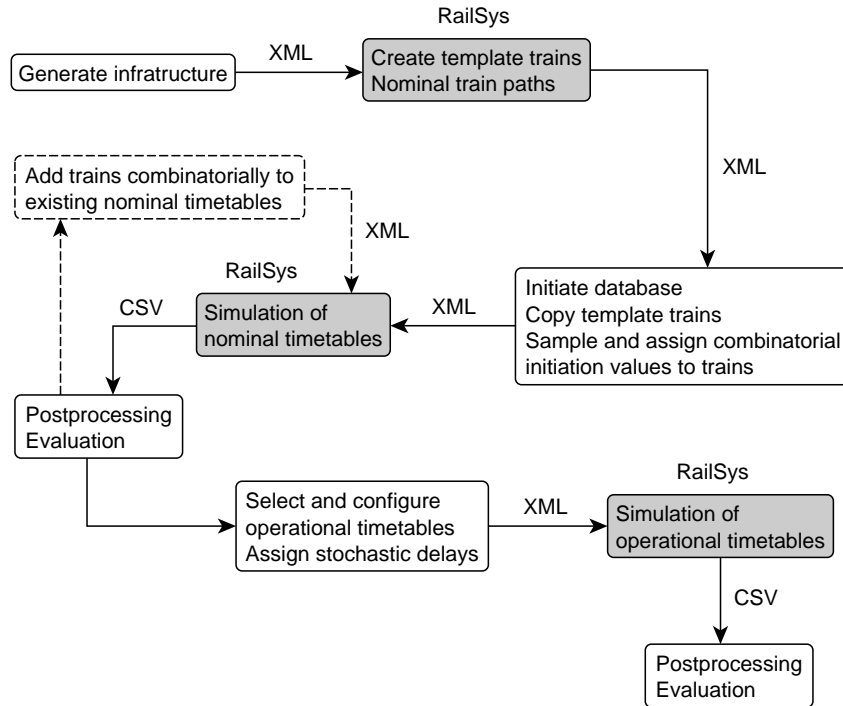


Figure 34: The process used for multiple timetable analysis with simulation of nominal and operational timetables. Data formats used for input and output in RailSys are indicated.

Multiple timetable analysis is useful since it counteracts some of the uncertainties inherited from making assumptions of future traffic in one or a few timetables. Static and dynamic performance measures can be analyzed and depending on how different properties are valued, a subset of candidate timetables can be chosen as input to other types of analyzes. By determining average and dispersion values for a subset, an infrastructure and traffic load variant can be represented and compared to other cases. Hence, the dots in fig. 23 are combined to one dot supplemented with a dispersion measure (fig. 35).

Instead of using a sufficient number of replications (cycles) in operational timetable simulations to obtain stability in results, fewer replications can be used for each timetable since they are aggregated to a group measure. The benefit is that simulation times can be decreased, but at the expense of the dispersion measure precision in the operational performance dimension. Despite which approach is used, the focus is lifted from analyzing specific timetables to the expected performance given a certain infrastructure and traffic load. Stochastic delays can be modeled based on assumptions and, if available, historical data from lines with similar conditions.

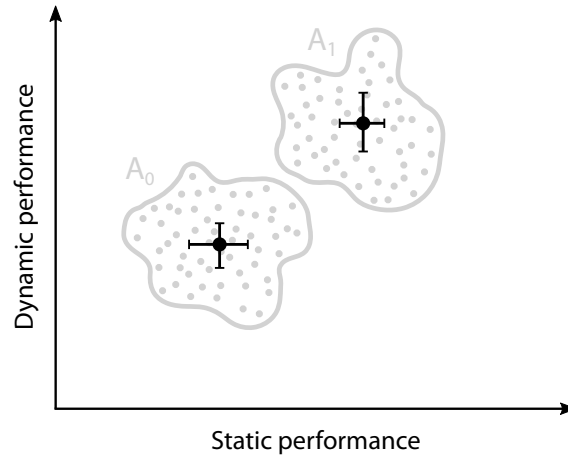


Figure 35: Combining subsets of timetables to an average value and dispersion measure.

The synchronous simulations performed to obtain operational timetables from an input of nominal timetables and later introducing stochastic delays to the operational timetables are performed in RailSys. External routines (scripts) were developed to create the combinatorial setup, handle postprocessing and evaluation of data as well as delay assignments and setting up the operational simulations. If necessary, the simulation setup can be distributed on multiple instances. A database is initiated prior to the combinatorial simulations. It is later filled with evaluation data in two steps, timetable and delay assignments for the operational simulations.

As previously indicated, the objective for developing the routines for infrastructure generation is to enable the creation of many ready-to-go infrastructure variants without manual intervention in RailSys and with small effort regarding time and needed repetitive operations. This gives also more control over which route options are needed and which are not without the need for corrections in the software. The possibility to generate multiple infrastructure variants facilitates the management of the infrastructure as a variable in capacity analysis of expansion alternatives in which timetables, operational delays, infrastructure layouts and ultimately costs are included to determine which solution(s) may be of interest.

8 Contribution of thesis

The contributions of this thesis can be divided in two parts. From a methodological point, refinement of simulation methods and the use of them in larger scale. From a practical point, to apply simulation models and construct timetables with the aim of reducing delays. As already mentioned on several occasions, the actual simulations are performed in RailSys. The methodological framework requires that many of the operations are made outside of RailSys in order to increase effectiveness and create appropriate input data and handle the analysis of output data.

8.1 Methodological framework

Micro-simulation of rail traffic requires a significant amount of input data and the setup is work and time intensive. Therefore, simulations are often restricted to one or a few specific timetable and/or infrastructure scenarios. Furthermore, it becomes strongly dependent on the individual making assumptions regarding the timetables. In this thesis, a methodology is developed which makes it easier to handle the infrastructure, timetables and delays as variables in a simulation study. This means that the results can be more objective and give a more comprehensive picture of how the variables interact in relation to the outcome.

This methodology uses a combinatorial approach for creating possible train departures and nominal timetables. Simulation is then used in two steps for obtaining operational (feasible) timetables and assess some of these when exposed to stochastic delays. The presented case applications concern essentially single-track lines, but the methodology is applicable for larger networks as well, although the very large number of combinations will require sampling. Even if the infrastructure can be defined manually, the infrastructure generator enables a considerably faster way of creating multiple alternatives.

The use of train registration data to compile distributions for deviations with respect to scheduled run times and systematically reducing the distributions provides an efficient way of estimating primary run time extensions. These are concluded to be important to model Swedish conditions, hence there is a need for a framework that can handle this. Another question during previous simulation studies has been how the limitation of not modeling early running freight trains affects the outcome.

The proposed methodology for capturing a more realistic variance in freight operations indicates that the effects on passenger trains are low. The finding is consistent with a Swedish field study from 2009. The methodology is still useful in studies where realistic freight train operations are important. In connection with this, improvements in the modeling of delays initiated by the handling of freight trains on stations can be made.

8.2 Practical framework

A simulation model was established for the Western Main Line in Sweden with the aim of evaluating the effect of specific timetable changes on the on-time performance for high-speed trains. The adopted changes consisted of decreased and increased allowances as well as increased buffer times. Many critical points where the buffer time requirement induced adjustments were identified. Although the aggregated on-time performance was slightly improved, some of the train individuals showed considerable improvement. Some of the findings were later adopted in the real timetable.

The work conducted within the Green Train project where data on curve radii was used to calculate possible speed profiles on both existing and future lines has mainly been used as input to other studies. It highlights the potential for faster run times that newer trains can have on existing lines with only minor changes to the track geometry. Some of the speed profiles and train configurations were adopted in a simulation study on the Southern Main Line which focused on the capacity when mixing fast passenger trains with other slower trains. The conclusion was that it is difficult to increase the top speed of high-speed trains from 200 to 250 km/h without losing capacity. On the other hand, it is possible to gain capacity if freight trains can increase their top speeds from 100 to 120 km/h.

8.3 Future work

An interesting application is to use more detailed data from a simulation to possibly analyze event chains, i.e. the detailed propagation of knock-on delays originating from a primary delay. Applied on a real timetable, it can improve the understanding of the relationship between these type of delays related to different train categories. The normal train registration data can explain some of these, but due to the lack of information such as track use and realized train arrival times on stations where a stop is not pre-scheduled, many assumptions would be needed for this type of analysis. Using detailed data from the interlocking systems can overcome this gap, but this type of data can be more challenging to handle from other perspectives and is not generally available as is the case with the basic registration data.

Another application deals with the calibration effort needed for larger simulations, especially concerning the modeling of run time extensions. It could potentially be useful to configure a library of template distributions for lines with different characteristics regarding the infrastructure, traffic load and train categories. Instead of template distributions, it may be a methodology framework for compiling these distributions out of registration data. The main reason is that it can potentially save a substantial amount of time in a calibration process and thereby facilitate the use of simulation in timetable planning and other studies.

References

- Abril, M., Barber, F., Ingolotti, L., Salido, M.A., Tormos, P., Lova, A., 2007. An assessment of railway capacity. *Transportation Research Part E* 44, 774–806.
- Anand, N., Anayi, M., 2009. Improving punctuality of train traffic on Western main line of Swedish railway network, in: *Proceedings of the AMSE 2009 Rail Transportation Division Fall Conference*, Fort Worth, USA.
- Andersson, E., Berg, M., Stichel, S., 2013a. *Rail Vehicle Dynamics*. School of Engineering Sciences, Aeronautical and Vehicle Engineering, KTH.
- Andersson, E., Persson, R., 2006. Fall för gångtidsberäkningar i Gröna tåget – Förslag (Run time calculation cases – Proposal). Memorandum, KTH. (In Swedish).
- Andersson, E.V., Peterson, A., Törnquist Krasemann, J., 2013b. Quantifying railway timetable robustness in critical points. *Journal of Rail Transport Planning & Management* 3, 95–110.
- Banverket, 2009. Testvecka: släpp inte ut tidiga tåg 2–6 mars (Test week: no clearing of early trains March 2–6). Swedish Rail Administration, Internal report. (In Swedish).
- Bendfeldt, J.P., Mohr, U., Müller, L., 2000. Railsys, a system to plan future railway needs, in: Brebbia, C., Allan, J., Hill, R., Sciutto, G., Sone, S. (Eds.), *Computers in Railways VII: Proceedings of CompRail 2000*, WIT Press, Bologna, Italy.
- Confessore, G., Liotta, G., Cicini, P., Rondione, F., Luca, P.D., 2009. A simulation-based approach for estimating the commercial capacity of railways, in: Rossetti, M., Hill, R., Johansson, B., Dunkin, A., Ingalls, R. (Eds.), *Proceedings of the 2009 Winter Simulation Conference*.
- Dagerholm, M., 2009. Metodapplikation i kvalitetsförbättring – kollektivtrafiken (Application method in quality improvement – public transport). Master’s thesis. Faculty of engineering (LTH), Lund University. (In Swedish).
- Dingler, M., Lai, Y.C., Barkan, C.P.L., 2009. Impact of train type heterogeneity on single-track railway capacity. *Transportation Research Record: Journal of the Transportation Research Board*, 41–9.
- Eliasson, J., Börjesson, M., 2014. On timetable assumptions in railway investment appraisal. *Transport Policy* 36, 118–26.
- 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, 2–13.

- Forsgren, M., Aronsson, M., Kreuger, P., Dahlberg, H., 2011. The Maraca – a tool for minimizing resource conflicts in a non-periodic railway timetable, in: Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis (RailRome), Rome, Italy.
- Fransoo, J., Bertrand, J., 2000. An aggregate estimation model for the evaluation of railroad passing constructions. *Transportation Research Part A* 34, 35–49.
- Fröidh, O., 2005. (Ed.) Gröna tåget – Framtida tågprestanda och bangeometri – Förstudie. Memorandum, KTH. (In Swedish).
- Fröidh, O., 2006. Gångtidsberäkningar i Gröna tåget (Run time calculations for the Green Train). Memorandum, KTH. (In Swedish).
- Fröidh, O., 2012. Green train – Basis for a Scandinavian high-speed train concept – Final report – Part A. Technical Report 12–01. KTH Railway Group.
- Gibson, S., Cooper, G., Ball, B., 2002. Developments in transport policy: The evolution of capacity charges on the UK rail network. *Journal of Transport Economics* 36, 341–54.
- Gille, A., Klemenz, M., Siefer, T., 2008. Applying multiscaling analysis to detect capacity resources in railway networks, in: Allan, J., Arias, E., Brebbia, C., Goodman, C., Rumsey, A., Sciutto, G., Tomii, N. (Eds.), *Computers in Railways XI: Proceedings of Comprail 2008*, WIT Press, Toledo, Spain.
- Hwang, C.C., Liu, J.R., 2010. A simulation model for estimating knock-on delay of Taiwan regional railway. *Journal of the Eastern Asia Society for Transportation Studies* 8.
- Kantamaa, V.M., Mäenpää, H., Pitkänen, J.P., 2013. Simulation as a planning tool for Helsinki railway station yard layout. *Signal + Draht* 10, 55–60.
- Koutsopoulos, H., Wang, Z., 2007. Simulation of urban rail operations: application framework. *Transportation Research Record: Journal of the Transportation Research Board* 2006, 84–91.
- Koutsopoulos, H., Wang, Z., 2011. Calibration of urban rail simulation models: A methodology using spsa algorithm, in: Jain, S., Creasey, R.R., Himmelspach, J., White, K.P., Fu, M. (Eds.), *Proceedings of the 2011 Winter Simulation Conference*, Phoenix, USA.
- Kufver, B., 2000. Optimisation of horizontal alignment for railways: Procedures involong evaluation of dynamic vehicle response. Ph.D. thesis. TRITA–FKT Report 2000:47, KTH.
- Kunimatsu, T., Sakaguchi, T., Ishihara, Y., 2013. Evaluation of facility improvements from the viewpoints of service level robustness for passengers, in: Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen).

- Labermeier, H., 2013. On the dynamic of primary and secondary delay, in: Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen), Copenhagen, Denmark.
- Lai, Y.C., Barkan, C.P.L., 2009. Enhanced parametric railway capacity evaluation tool. *Transportation Research Record: Journal of the Transportation Research Board*, 33–40.
- Lai, Y.C., Barkan, C.P.L., 2011. Comprehensive decision support framework for strategic railway capacity planning. *Journal of Transportation Engineering* 137, 738–49.
- Landex, A., 2010. Computation and evaluation of scheduled waiting time for railway networks, in: Ning, N., Brebbia, C. (Eds.), *Computers in Railways XII: Proceedings of CompRail 2010*, WIT Press, Beijing, China.
- Lindahl, A., 2002. *Infrastruktur för flexibel tågkörning: Kapacitetsanalys av förbigångar på en dubbelspårssträcka* (Infrastructure for flexible train operation: Capacity analysis of overtakings on a double-track railway line). Technical Report TRITA-INFRA 02-021. KTH. (In Swedish).
- Lindfeldt, A., 2009. *Kapacitetsanalys av järnvägsnätet i Sverige* (Capacity analysis of the rail network in Sweden). Technical Report TRITA-TEC-RR 10-003. KTH. (In Swedish).
- Lindfeldt, A., 2014. *Kapacitetsutnyttjande i det svenska järnvägsnätet: Uppdatering och analys av utvecklingen 2008–2012* (Capacity utilization of the Swedish rail network: Update and analysis of the development 2008–2012). Technical Report TRITA-TEC-RR 14-003. KTH. (In Swedish).
- Lindfeldt, O., 2011a. Analysis of capacity on single-track railway lines, in: Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis (RailRome), Rome, Italy.
- Lindfeldt, O., 2011b. An analysis of double-track railway line capacity. *Transportation Planning and Technology* 34, 301–22.
- Lindfeldt, O., Sipilä, H., 2009. Validation of a simulation model for mixed traffic on Swedish double-track railway line, in: Forde, M. (Ed.), *Proceedings of Railway Engineering 2009, 10th International Conference and Exhibition*, Engineering Technics Press, London, UK.
- Marinov, M., Şahlin, İ., Ricci, S., Vasic-Franklin, G., 2013. Railway operations time-tabling and control. *Research in Transportation Economics* 41, 59–75.
- Marinov, M., Viegas, J., 2011. A mesoscopic simulation modelling methodology for analyzing and evaluating freight train operations in a rail network. *Simulation Modelling Practice and Theory* 19, 516–39.

- Mattsson, L.G., 2007. Railway capacity and train delay relationships, in: Murray, A., Grubescic, T.H. (Eds.), *Critical Infrastructure: Reliability and Vulnerability*. Springer-Verlag. chapter 7, pp. 129–50.
- Milinković, S., Vesković, S., Marković, M., 2013. Modelling train delays in rail networks with large disturbances, in: *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen)*, Copenhagen, Denmark.
- Nash, A., Huerlimann, D., 2004. Railroad simulation using OpenTrack, in: Allan, J., Brebbia, C.A., Hill, R.J., Sciutto, G., Sone, S. (Eds.), *Computers in Railways IX: Proceedings of CompRail 2004*, WIT Press, Dresden, Germany. pp. 45–54.
- Nelldal, B.L., Lindfeldt, O., Sipilä, H., Wolfmaier, J., 2008. Förbättrad punktlighet på X2000: Analys med hjälp av simulering (Improving punctuality for X2000: Simulation analysis). Technical Report TRITA-TEC-RR 08-001. KTH. (In Swedish).
- Olsson, N., Haugland, H., 2004. Influencing factors on train punctuality – results from some Norwegian studies. *Transport Policy* 11, 387–97.
- Pachl, J., 2002. *Railway operation and control*. VTD Rail Publishing, Mountlake Terrace, WA, USA.
- Pachl, J., 2007. Avoiding deadlocks in synchronous railway simulations, in: *Proceedings of the 2nd International Seminar on Railway Operations Modelling and Analysis (RailHannover)*, Hannover, Germany.
- Pachl, J., 2011. Deadlock avoidance in railroad operations simulation, in: *Transportation Research Board. 90th Annual Meeting*.
- Persson, R., 2010. Tilting trains – benefits and motion sickness. *Journal of Rail and Rapid Transit* 224.
- Peterson, A., 2012. Towards a robust traffic timetable for the Swedish Southern Mainline, in: Brebbia, C., Tomii, N., Mera, J., Ning, B., Tzieropoulos, P. (Eds.), *Computers in Railways XIII: Proceedings of CompRail 2012*, WIT Press, Lyndhurst, UK. pp. 473–84.
- Pouryousef, H., Lautala, P., 2014. Evaluating two capacity simulation tools on shared-use U.S. rail corridor, in: *Transportation Research Board. 93rd Annual Meeting*.
- Quaglietta, E., Punzo, V., Montella, B., Nardone, R., Mazzocca, N., 2011. Towards a hybrid mesoscopic-microscopic railway simulation model, in: *Proceedings of the 2nd International Conference on Models and Technologies for Intelligent Transportation Systems*, Leuven, Belgium.
- Radtke, A., 2008. Infrastructure modelling, in: Hansen, I., Pachl, J. (Eds.), *Railway Timetable & Traffic*. Eurailpress. chapter 3, pp. 43–57.

- RailSys, 2014. Version 9 user manual. Rail Management Consultants (RMCon). URL: www.rmcon.de.
- Rudolph, R., 2003. Allowances and margins in railway scheduling, in: Proceedings of the 6th World Congress on Railway Research, Edinburgh, Scotland. pp. 230–8.
- Siefer, T., 2008. Simulation, in: Hansen, I., Pachl, J. (Eds.), *Railway Timetable & Traffic*. Eurailpress. chapter 9, pp. 155–69.
- Sipilä, H., 2008. Körtidsberäkningar för Gröna tåget: Analys av tågkonfigurationer (Run time calculations for the Green Train: Train configuration analysis). Technical Report 08–02. KTH Railway Group. (In Swedish).
- Sipilä, H., 2010. Tidtabellsläggning med hjälp av simulering: Effekter av olika tillägg och marginaler på X2000-tågen Stockholm–Göteborg (Timetable planning using simulation: Effects of supplements and allowances for X2000 trains Stockholm–Gothenburg). Technical Report TRITA–TEC–RR 09–007. KTH.
- Sogin, S., Barkan, C.P.L., Saat, M.R., 2011. Simulating the effects of higher speed passenger trains in single-track freight networks, in: Jain, S., Creasey, R.R., Himmelpach, J., White, K.P., Fu, M. (Eds.), *Proceedings of the 2011 Winter Simulation Conference*, Phoenix, USA.
- Sogin, S.L., Lai, Y.C., Dick, C.T., Barkan, C.P.L., 2013. Analyzing the incremental transition from single to double track railway lines, in: *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen)*, Copenhagen, Denmark.
- Trafikanalys, 2015. Järnvägstransporter 2014, kvartal 4 (Railway transport 2014, quarter 4). Transport analysis. (In Swedish).
- UNECE, 2014. Number of railway passengers by country, passengers and time. United Nations Economic Commission for Europe, statistical database.
- Warg, J., Bohlin, M., 2015. The use of railway simulation as an input to economic assessment, in: *Proceedings of the 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo)*, Tokyo, Japan.
- White, T., 2005. Alternatives for railroad traffic simulation analysis. *Transportation Research Record: Journal of the Transportation Research Board* 1916, 34–41.
- Östlund, S., 2012. *Electric Railway Traction*. School of Electrical Engineering, Electrical Energy Conversion, KTH.

