Postprint

This is the accepted version of a paper presented at *Railway Engineering 2013 12th International Conference and Exhibition; London, UK 10–11 July, 2013.*

Timetable generation on single-track lines using combinatorics and simulation

Hans Sipilä

KTH Royal Institute of Technology

Department of Transport Science

100 44 Stockholm, Sweden

hans.sipila@abe.kth.se

Abstract

Timetable construction on lines with dense and heterogeneous traffic can be demanding. Mixing freight and passenger trains on single-track lines adds complexity in scheduling. Simulation offers a way of estimating operational outcome prior to establishing the actual timetable. Realistic modelling of delays is important in simulations. Exogenous delays can cause knock-on effects between trains, i.e. delay propagation.

Many scheduling approaches do not study the expected operational outcome, similarly simulation studies are mostly done on a small number of timetables. This paper presents a method that combines generation and simulation of timetables by using a combinatorial approach as input data to a simulation software and letting it compute corresponding solutions. Measures of performance in choosing timetables for further studies can be amount of scheduled delay, requirements on cyclic patterns etc. In the next step chosen timetables can be simulated with perturbations and give insight of expected operational performance. This can in turn be used to give a more general assessment of possible capacity on a line.

The method is applied on a fictive single track line but the same principles could also be used on double tracks. Simulation software RailSys is used in this study, among its users are for example operators, infrastructure managers and universities. RailSys offers no built in method for generating a large number of different timetables but it includes a powerful simulation module.

1 Introduction

Increasing demand for rail transportation often implies a higher track utilization rate. Dealing with this increase in number of passengers and tons of freight typically means that more trains are operated. Other ways of improving rail transport capacity are for example to use longer and/or heavier trains. Typically a mix of measures is used to meet the transport demand. However, the existing infrastructure imposes limits on possible actions. Low speeds, imposed by the track itself or by trains, leads to longer run times between locations where passing and meeting of trains is possible. Inter-station distances contribute further to temporal conditions, particularly on single track lines. Number of tracks and track lengths on stations constrains how many and which train configurations can be handled simultaneously. The signalling system influences for instance headways between trains, maximal axle and meter loads relate directly to train weights and so on. Consequently, the capacity of a railway line is defined by several parameters with and without inter-dependencies.

1.1 Timetable factors

Most railways are operated on a timetable basis, meaning that train runs are scheduled in advance to be more or less conflict free. Deviations from schedule will always occur to a degree, depending on the type and mix of operated services. On-time performance requirements for passenger trains are usually high and the feedback response from travellers is often quick when delays are frequent and high. Requirements on freight services are high as well, but partly due to the operational structure with freight cars using different trains and passing marshalling yards between their origin and final destination, there can be substantial slack in a circulation schedule enabling delay reduction before a freight car reaches its customer. Although not true in general, it is not uncommon that freight trains have a lower dispatching priority compared with many passenger trains. This can be seen both in the planning and operational phase and leads to waiting times freight trains when passed by or meeting faster and more prioritized trains.

Small deviations from schedule should be absorbed by allowances combined with buffer times. Allowance or margin time is often described as the difference between minimum and scheduled run time on some considered distance, i.e. there is some extra time which can be useful in reducing or avoiding delay propagation. The definition of minimum run time varies, e.g. it can be a sum of technical minimum run time and some percentage standard allowance. This minimum time is supplemented with additional allowance that can for example be distance based, i.e. longer running distance leads to more total allowance. It can also occur as a speed equalization measure on heavily utilized sections with a mix of train characteristics and stopping patterns. Buffer times are scheduled headways (time spacing) between trains, commonly illustrated with two consecutive trains running in the same direction on a track. However, buffer times are of interest in all locations where a smaller delay on one or more trains implies a route conflict.

Although margin and buffer times consume capacity 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 perturbations leading to delay propagation. In this context a categorization is made into primary and secondary delays in which the first one refers to exogenous events affecting trains and the second one to knock-on effects, i.e. delays caused by interaction between trains. Considering a railway system with a conflict free timetable, meaning that there are no route conflicts caused by overlapping occupation times, all knock-on events originate from one or several exogenous events. The chain of events leading to a specific knock-on event is often hard to distinguish in a system with many trains and dense traffic.

A timetable that is achievable is influenced by several factors which in turn have subfactors and so on. Some of these are exemplified in figure 1. For example, locations of single- and double-track sections, points and signalling and so on fall under the infrastructure factor. Available vehicles and their performance give possible trip chaining schedules which can also be affected by staff working time rules etc. Experience and analysis of observed outcome give input for allocations of margin 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.
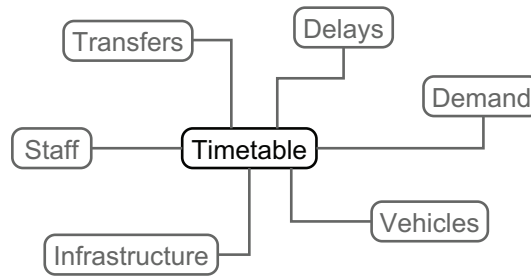
Figure 1: Examples of factors influencing the timetable construction process.

1.2 Rail traffic simulation

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 estimate the impact on measure of performance (MOP) parameters when changes are made to timetables, infrastructure, frequencies and levels of primary perturbations and so on. Many actions that are taken, e.g. infrastructure expansions or improving vehicle performance, offer the possibility of changing the timetable as well. Smaller adjustments serve a purpose of adding more flexibility to real operations whereas larger measures can enable significant timetable changes.

Infrastructure can be modelled on macro- or microscopic level depending on the effects a specific study aims to capture. In both cases nodes and connecting links define the track layout, i.e. the network. Briefly described, a microscopic approach model switches, signals, distant signals, speed boards, gradients, stopping points on stations and other objects considered necessary in a simulation. Macroscopic models imply a more aggregated description, stations can for example be defined by single nodes with some characteristics instead of detailed microscopic track layouts made up by several nodes and links.

Mesoscopic models fall between the previously mentioned, enabling more detail in some areas or for some features while keeping the macroscopic description in other parts. A full scale microscopic simulation approach in a large network is data and work intensive but has the potential of giving results with high accuracy. However, macroscopic modelling is eligible for high-level tasks such as strategic decision making on traffic demand and allocation.

Train run times, including acceleration and braking characteristics, are needed for designing timetables on a functioning infrastructure model. In this stage occupational conflicts between trains can be dealt with, before proceeding to simulations. The stochastic behaviour of a railway system is emulated by introducing perturbations reflecting variations in passenger exchange times, run times between stations, train initiation in the network and so on. This can both be used to model regularly occurring smaller variations and systematic disturbances.

Different kinds of perturbations appear with varying probabilities and magnitudes. Therefore, the probability of a disturbance 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.

Assuming that primary delays are modelled reflecting real conditions simulations with different setups give outcome that can in turn be compared with respect to each other. If a base scenario exists, this can be used for comparisons. Typically a base scenario can model the current situation, typically today's timetable. However, if the timetable is not given beforehand an obvious problem or challenge is to find or design one or several sufficiently good timetables forming the simulation cases. Given some prerequisites, there should at least be a perception of what characterizes a good timetable and by that facilitate ranking among the studied alternatives.

2 Related research

In Radtke (2004) and Siefer (2008) different simulation methods are described, e.g. synchronous and asynchronous modelling. Differences between microscopic and macroscopic approaches are also explained. Furthermore, the basic functionalities of RailSys are presented. Allowances and margins in railway scheduling is discussed in Rudolph (2003). Recovery times can decrease or absorb delays for specific trains. These are made up of time supplements added to the minimum run time. Buffer times are added to minimum headway between two trains and act as recovery times of the timetable. The objective with these properties is to increase the reliability or robustness of timetables, i.e. to improve their capability of handling smaller regularly occurring delays.

Capacity on single track lines with additions of both freight and passenger trains is investigated in Sogin (2011). Simulations on a representative fictive route with varying train density are performed with Rail Traffic Controller (RTC). A homogeneous condition with a composition of 100% freight trains is defined as a base case, although the number of trains varies. Comparing 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 the delay of the freight trains. Higher passenger speeds will increase delays although the marginal increase in speed has a declining factor on the delays of freight trains when the network becomes saturated.

A hypothetical single-track line is modelled in Dingler (2009). 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 measures 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. Due to both higher priority and speeds, adding passenger trains gives higher delays compared to adding the same number of freight trains.

Lindfeldt, O. (2011) presents a timetable evaluation method for single track lines (TVEMS). It is a purely deterministic capacity model thait aims to model the timetables of a single track line. A combinatorial approach is used to generate timetable variants for periodically operated passenger traffic by systematically time shift patterns in relation to each other. Non-periodical freight trains are modelled with predefined and randomised scheduling orders. Trains are introduced in sequence, starting with the highest prioritised train group one direction first and thereafter continuing with the same group in the other direction. A higher prioritised train is never given a scheduled delay through a meet or overtaking.

A field study of the Red Line in Boston is presented together with a simulation model (SimMETRO) in Koutsopoulos (2006). 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. The calibration process can be modelled as a multi-variate optimisation problem and solved with SPSA algorithm (Koutsopoulus, 2011). This approach is used in the simulation study discussed in Koutsopoulos (2006) and results show that the calibration process improves the parameters and refines the input.

A study of relationships between heterogeneity, train density, inter-station distance and secondary delays shows that trains with low priority are more likely to experience delays on stations, while higher prioritised trains suffer from extended run times between stations (Lindfeldt, A., 2011). A fictive double track line is modelled with overtaking stations and several traffic mixes are simulated. Both high speed and regional passenger trains are considered in combination with freight trains. Compared to homogenous timetables, the consumed allowance and secondary delays increase faster in heterogenous cases.

Some of the factors influencing train punctuality is discussed in Olsson (2004). Examples of studied factors are infrastructure capacity utilisation, 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 travellers, capacity utilisation 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.

Some topics regarding railway capacity and delay relationships are presented in Mattsson (2007). Capacity variation due to speeds, number of tracks etc. is illustrated with examples. Analytic, statistical and micro-simulation methods for delay analysis are presented and some of their advantages and drawbacks are listed. Analytic methods can be useful at a strategic planning and design stage, however simplifying assumptions are often necessary. This may not be a problem considering that many future parameters are uncertain. Elements from queuing theory and optimisation are common in these approaches. Micro-simulation offers a useful approach in modelling interactions between trains, i.e. propagation of knock-on delays initiated by some exogenous event.

A method for finding the level of primary run time extensions from train registration data is described in Sipilä (2011). Delay distribution are compiled for several train groups running on the Swedish Southern Main Line. These distributions contain both primary and secondary (knock-on) delays and the objective is to reduce them in several steps and use simulation to find good enough levels. The root mean square error is applied when comparing results from different setups. Distribution levels were varied in four steps from 20 to 80% and in this case best overall fit was achieved by using a 60% level for passenger trains and 40% level for freight trains.

3   Prediction and performance measures in rail traffic

Studying historical data from actual railway operations give valuable insight and can help in understanding which measures to undertake in order to improve future performance. Being able to predict, at least to some degree, how a timetable will perform is important since it is often difficult to attain changes once a timetable becomes definite or is running. This section outlines some timetable describing properties and typical measures used in evaluation rail traffic performance.

3.1   Timetable characteristics

There are also ways of assessing a timetable itself and predict possible outcome. Some measures are based on viewpoints that relate more to how attractive train slots and total run times are to operators and, in the end, to customers. One example of this is the scheduled delay, meaning the difference between an undisturbed run time and what can be accomplished in a timetable considering other trains and their accompanying demands as well. Some level of scheduled delay is preferable seeing that it can be used for delay recovery, thus it relates to some expected probability of disturbances.

Heterogeneous traffic, meaning that there is a mix of trains with different run times (average speeds), is often prone to cause delay propagation compared to a homogeneous timetable. Faster trains catching up slower trains due to disturbances call for dispatching actions to shift the train order and counteract further delay propagation for the faster train. On the other hand, the slower train may suffer from a delay and ultimately cause problems in some other location. Measures of heterogeneity exist and these can be used to give an estimation of the outcome in real operation.

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 margins, headways (buffer times) and heterogeneity.

The frequency of passings affects capacity and is closely linked to delay sensitivity. This holds for single track meetings 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 which may reduce delays in cases in which these type of stops are omitted due to the operational situation.

In heavily utilized networks trade-offs are needed and they affect the approved timetable. Requests for regular interval train slots are met to a varying degree. Although this contributes in making train travel attractive, it is not always the best solution from a capacity utilization viewpoint. At least smaller adjustments to regular interval systems should be contemplated if it can lead to a better overall solution.

3.2 Measures of performance

Simulations with applied perturbations aiming to create stochastic or systematic events are often evaluated with respect to a reference timetable. This can be done train by train or as aggregated values for several trains in unifying groups. Which measure of performance (MOP) is preferable depends on what effects are studied, type and mix of traffic, etc. Resulting effects of delayed trains as missed connections, number of affected passengers, socio-economic costs etc. are also important in a wider perspective.

One common MOP is on-time performance, also referred to as punctuality. It gives the percentage of trains up to a defined delay threshold at some specific measurement point in the network. On-time performance is often aggregated for many train groups and can for example refer to terminus arrival values. Typical delay thresholds vary between different networks, countries and perhaps most importantly on the type of operations.

Studying on-time performance en route can highlight sections with substantial changes in this measure, thus indicating both on worsening and improving delay situations. This is useful in timetable planning and gives some guidance for implementing adjustments in order to improve conditions. Although on-time performance is easily interpreted it does not reveal any further information about performance for trains that are considered on time and those falling outside threshold values. Additional measures and more detailed studies are needed to better capture the situation.

Taking the average of deviation values basically gives a delay measure which can be compared to threshold values and categorized according to a low-medium-high scale or some other reference. Mean values are sensitive to extremes (outliers) relative to the majority of deviation values, e.g. mean delay 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 true value is obtained if a cut-off is done at some level meaning that extremes are removed.

Adding 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. If at least 50% of the trains in a evaluation group have no registered delay, i.e. deviation is either zero or positive (ahead of schedule), the median value is often zero and therefore does not give much information about delays in the second half of the group. However, if the median is used only for delayed trains it becomes more interesting combined with mean and standard deviation.

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 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.

4 Initiation of start sequences

Designing timetables is sometimes straightforward due to a small network size, limited number of train systems (patterns) and conditions which impose temporal and spatial dependencies between trains. However, with increasing number of trains and larger networks it is apparent that there will often exist many possible timetable solutions. Depending on how the solutions are valued, some will most likely stand out as better compared to others. Timetable evaluations in itself calls for a number of different alternatives on which to carry out simulations or some other analysis method. On the other hand, the timetable can be kept constant while some other conditions are varied.

This section presents a method for feeding an existing rail traffic simulation software (RailSys) with systematic train initiation times and use the integrated dispatching function to produce timetable solutions. The aim is to see if this approach provides meaningful results and be of use in a future study of an existing Swedish railway line. This is exemplified on a case with a single track line and crossing stations. The same principles apply to double track lines or networks with several lines and junctions. Limitations are imposed by the number of sequences in combination with network size and number of trains that are, in a practical sense, possible to simulate regarding computational time.

Timetable input values are compiled from initiation of train start sequences. This means that every train is given a requested departure time from their respective first station and this time is varied following a combinatorial approach. Given a set of preconditions regarding number of train groups and their density (interval), all possible combinations can be obtained. However, for this to work two key parameters are needed:

- Time resolution (T_R)
- Sequence time (S_T)

The first one sets a minimum time between train initiations and also restricts possible initiation time slots. Sequence time can be seen as the total cycle time defined by the train groups and their density. Table 1 exemplifies number of combinations for some different setups with a sequence time of one hour. The number of combinations grows quickly with increasing time resolution and number of train groups. This becomes even more apparent when sequences for two directions are combined on a single track. Therefore measures are needed to limit the number of combinations prior to simulations, i.e. make a selection based on typical constraints set by the infrastructure and knowledge from real operations.

In reality, headway rules or policies are taken into account in a timetabling process. The first line block section provides a restriction for the headway time between two successive trains exiting a station; the first train must clear the block section before the following train can proceed. If a slower train precedes a faster, the full distance and block lengths between stations becomes dimensioning assuming a passing event at the second station. In the inverse situation, where a faster train leaves first, the first block section length indirectly provides a minimum headway time. Reasoning in this way assumes that block length are approximately equal. When real timetables are constructed these values are typically standardized for some template trains and can include additional margin.

Table 1: Number of combinations for different time resolutions with one departure per hour and direction for each group.

Train groups	Direction	Time resolution (min)		
		5	3	1
2	1st	11	19	59
	2nd	132	380	3540
	Total	1452	7220	$2.09 \cdot 10^5$
3	1st	110	342	3422
	2nd	1320	6840	$2.05 \cdot 10^5$
	Total	$1.45 \cdot 10^5$	$2.34 \cdot 10^6$	$7.03 \cdot 10^8$

The time resolution parameter (T_R) defines the initiation times that can be used within the sequence time. Available time slots are given by $n \cdot T_R$ where $n = 0,1,\ldots,(S_T - T_R)/T_R$. The minimum temporal distance between train initiations is given by a combination of T_R and defined headway parameters between train groups. If minimum headway between a train in group j following a train in group i is H_{ij}, then the minimum temporal distance $(T_{\min})$ that will be used is given by

$$T_{\min \atop ij} = \left\lceil \frac{H_{ij}}{T_R} \right\rceil \cdot T_R \tag{1}$$

Figure 2 illustrates different sequences for a setup with three train groups with one departure per hour. Time resolution and minimum headway is three minutes except for trains following train group 3 which have a nine minute minimum headway. This is a typical case where trains in one group have a considerably longer run time compared to the other groups. There is no need to vary train group 1 since this would yield time shift repetitions of already obtained sequences. If figure 2 is considered a subset of a cyclic timetable, the same sequences are extended over several hours.
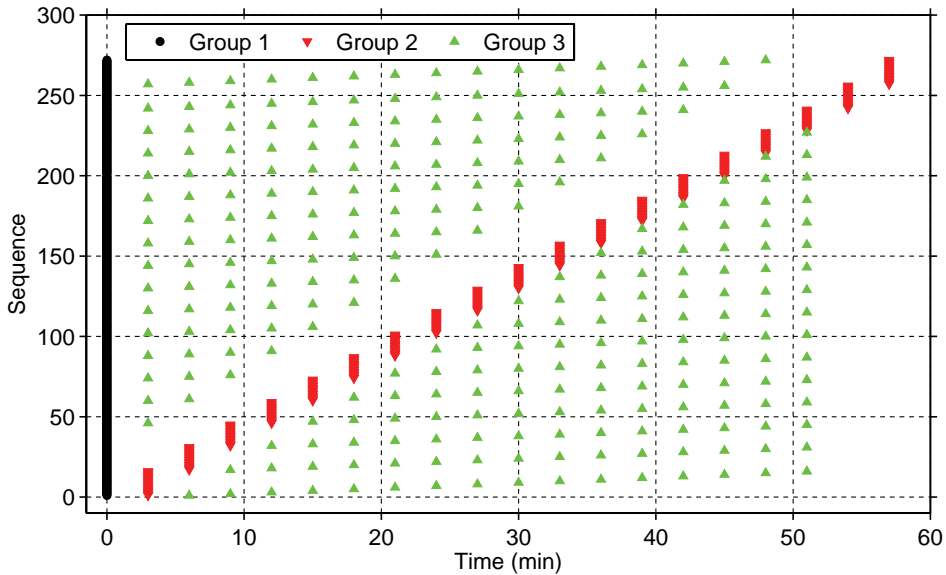

Figure 2: Example of possible sequences with three train groups, three minute time resolution and deviating headway rule for trains following a group 3 train compared with the other cases.

Considering a single track line, this case gives 5440 sequences in the opposite direction. The difference in numbers between directions comes from the fact that the first train group cannot be fixed on a specific departure time since this would not describe all possible sequences when both directions are combined. The total number of sequences comes close to 1.5 million and this calls for measures that can reduce this number to make it manageable for simulation. This combinatorial effect is also clearly seen in table 1, adding a fourth train group would increase these numbers even further.

One method that will reduce the problem is to consider which stations are desirable for meetings between trains in the same group or with respect to another group. Trains having scheduled stops for passenger exchange at certain stations will gain time if a meeting is done simultaneously instead of carrying it out at stations without passenger stops. This assumes that waiting times do not exceed the limit at which a meeting should be moved to an adjacent station in order to minimize total scheduled delay or some other dimensioning measure of performance. In the sequence generation process this means that one or more train groups are fixed to some slots defined by the time resolution and vary over these instead of the full sequence time.

Figure 3 gives an example of an adopted limitation of initiation times for group 1 trains, here only time slots 16, 18 and 20 are considered. If this represents trains in the opposing direction, i.e. the number of sequences is multiplied with the first direction, a significant drop in the total number of sequences is achieved.
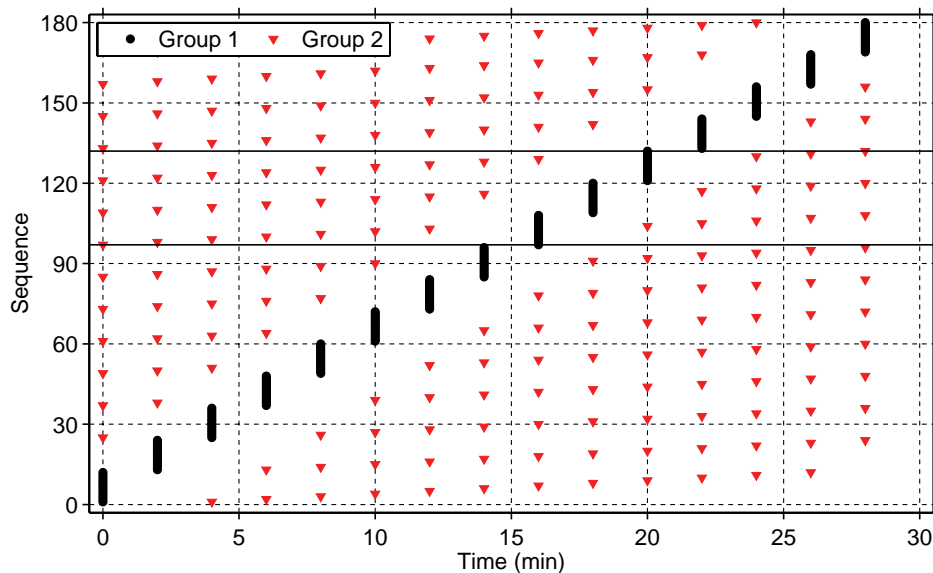

Figure 3: Example of situation where the number of considered sequences is limited to the range defined by the solid horizontal lines.

The procedure in figure 3 is partially based on the assumption that a meet can be estimated to occur at a certain station for trains in group 1. However, this is not known in advance since the complete timetable for a specific sequence is obtained by combining this input and simulations in RailSys. Estimations can be assumed to have a good enough accuracy if the train group in question has a higher dispatching priority than other groups in the system. Another approach that can be used is to compute and simulate sequences for two highest ranked groups first and find a subset of timetables with respect to one or more measures of performance values and compute sequences in a second step with additional train groups and the existing groups fixed. Applications of this are discussed further in section 5.

5   Case application

The approach with sequence initiation is applied on a single track line and simulations are carried out in RailSys. The modelled infrastructure is fictive but it shares some features with Swedish signalling and ATC. The track layout consists of eleven stations, with one passing siding, where meetings and overtakings can take place and these are placed nearly equidistantly with 10–11 km between their midpoints, no track gradients are used. Total running distance for trains from first to last station is close to 106.2 km.

Maximum speed is 200 km/h throughout the whole line, diverging speed in switches (turnouts) is 100 km/h. The line between stations is divided in two equidistant block sections ($\sim$ 5 km), i.e. there can at most be two trains running in the same direction on the line between successive stations. Target point information is only updated at main signal and repeater distant signal balises which are located 1000 and 300 m prior to entrance signals. This means that transmission of signal information to trains is non-continuous. If the target speed information is expect stop, the release speed at which deceleration supervision ceases is 40 km/h. Target points for side track speeds are extended from station entry signals towards switches.

Train characteristics represent three configurations with acceleration performance according to figure 4. The objective is to model mixed traffic conditions with high speed (InterCity), regional and freight services with top speeds 200, 160 and 100 km/h. Dispatching priority is set with respect to top speed meaning that delay increases for high speed trains are viewed as having a higher cost than for trains with lower priority. However, this parameter can be varied hence giving different weights regarding delay increase for a higher prioritized train and total delay increase for both trains involved in a route conflict.
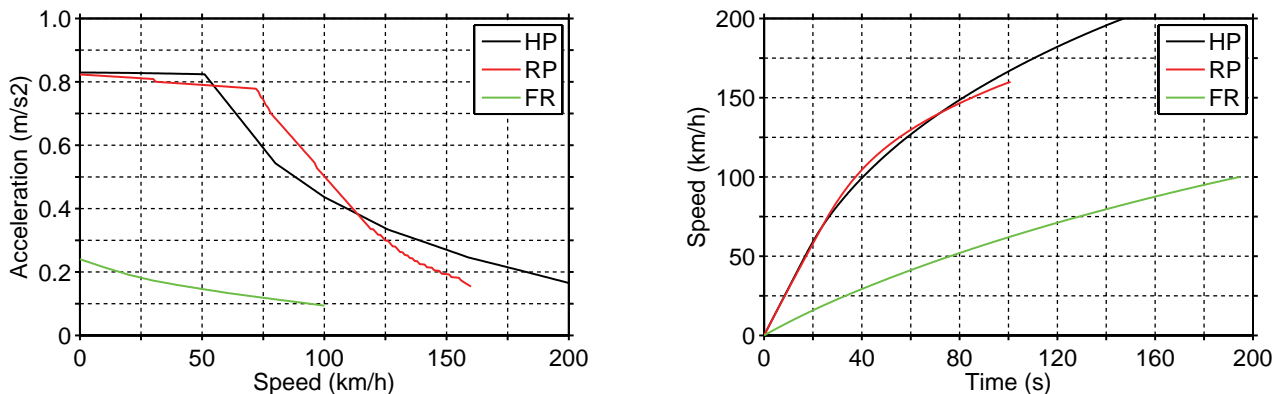

Figure 4: Flat track acceleration characteristics for high speed passenger (HP), regional passenger (RP) and freight trains (FR). Braking is modelled with constant deceleration rates, 0.6 m/s^2 for passenger trains and 0.3 m/s^2 for freight trains.

All trains start with an acceleration from standstill at the first station and end with deceleration to standstill at the last station. Table 2 shows reference run times, dwell and regularity. A standard allowance of 3% is applied and locked in the simulations, this means that the technical minimum run time is slightly lower than table values. Simulations are mainly evaluated with respect to scheduled delay, i.e. the difference between undisturbed and actual run time.

Simulations are prepared according to methods described in section 4. Technically speaking, the sequences are modelled by giving specific trains an initial delay. Knowing the requested and actual defined departure in RailSys gives the initial delay by taking the difference between these two times. Initially, a setup with two passenger train groups is tested. Results from this run are used in a second step in with added freight trains. Also a scenario with improved track capacity on stations

is introduced to see how this will affect the results with respect to the corresponding case with two track layout. In the following sections the term train group refers to table 2 with both directions combined.

Table 2: Reference run times including scheduled dwell per stop and regularity in start sequence

Train	Time	Scheduled stops	Regularity
High speed (HP)	0:39	One stop, 2 min	60 min
Regional (RP)	0:49	Three stops, 1 min	60 min
Freight (FR)	1:08	No stops	120 min

5.1 Case with passenger trains

Passenger train groups are initiated with regularity according to table 2, i.e. one train per group and per direction every hour. Time resolution is set to $T_R = 3$ min. Train group HP is fixed on the first time slot as illustrated in figure 2 for one direction. The other groups, RP in both directions and HP in the opposite direction, are floating. In this case a minimum headway (H) of six minutes is defined giving that $T_{\min} = 6$ min as well due to the choice of T_R.

The main parameter used in result evaluation is scheduled delay, for one or several groups combined. Some measures of regularity at last station are introduced to determine if and how these correlate with scheduled delay. Sequence time (S_T) is 60 minutes and this is repeated 10 times to get sufficient number of train runs for evaluation. During processing of simulation data, the first and last sequences are removed since they are considered warm up and cool down periods in every simulated cycle and do not represent full traffic. The reported input values generate 5780 start sequences in total. This number decreases to 5698 during the simulation process, the causes are discussed later.

Figure 5 shows the group mean of absolute scheduled delay and the mean percentage of increased run time with respect to table 2. Actual run time in simulations is based on the initiation time, which can be seen as a requested departure time, and arrival at last station in including dwell en route. In other words, even if trains are not able to depart on their initiation time, e.g. due to that the first line section is occupied by another train, this waiting time contributes to the considered run time.
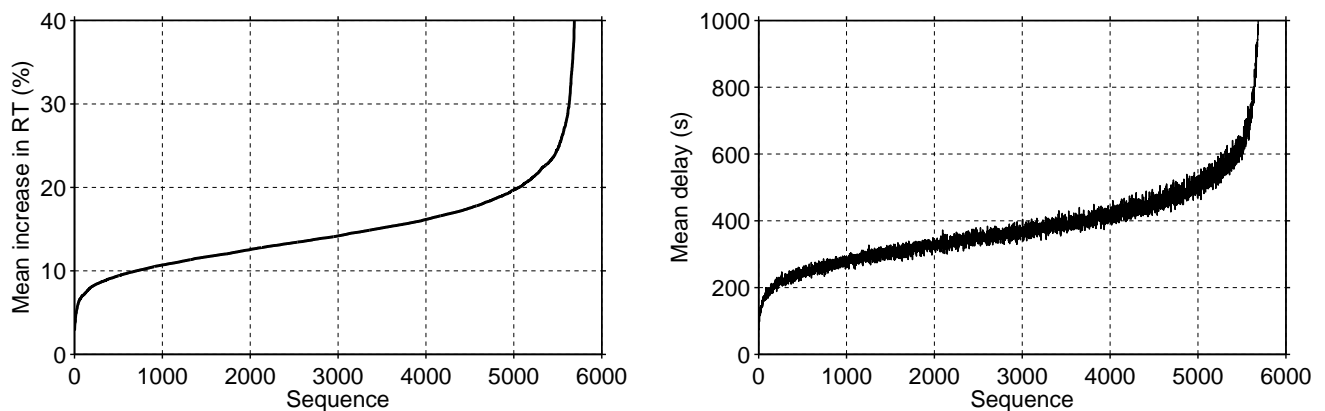

Figure 5: Mean percentage of increased run time (RT) and mean scheduled delay for all groups. Sorting from lowest to highest value with respect to mean percentage of increased run time.

One way of interpreting figure 5 is that there are relatively few sequences with distinguished low delay values. It can also be seen that the two measures in figure 5 does not completely coincide when both are sorted based on one of the measures. Most sequences fall in the range of 10–20% or 250–500 seconds. Since every train will experience meets with other, trains the existence of a sequence giving no delays is unlikely. In theory this is possible if all meets end up on stations with scheduled stops and sync perfectly with respect to dwell time. A small number of sequences results in high delays, these are typically caused by situations where the simulation software dispatching is faced with a situation where the first decided solution cannot be carried out due to changed conditions induced by opposing traffic.

Assessing the cyclicity of a timetable or selected train groups can be of interest in choosing timetables (i.e. sequences) for further processing. Even though some deviation is acceptable it should not be too large, hence making the timetable unattractive from a passenger view point. The term cyclicity is here used as a measure of the variation in arrival lateness at final station per train group. In other words, the regularity in temporal distance between arrival times.

In figure 6 cyclicity ia approached in two different ways. The first measure considers the number of unique run times per train group, or equivalently using lateness values. To deal with small deviations, values are floored to a resolution of 120 seconds. If N represents the number of trains in evaluation group j and n_i is the number of unique run times in sequence i, the cyclicity (C_{ij}) is given by

$$C_{ij} = \frac{n_i - 1}{N - 1} \qquad (2)$$

All values equal in a sequence gives $C_{ij} = 0$, analogously $C_{ij} = 1$ if all values are unique. It should be clarified that this measure requires equal number of trains in the groups on which it is applied. Here all four groups have six trains remaining after removal in beginning and end of the simulation period where full traffic is not expected. Total cyclicity is computed with equation 2 and by taking the mean for all groups in each sequence.

If this value is sorted with respect to mean percentage increase in run time according to figure 5, a highly varying curve is obtained. In figure 6, a central moving mean gives a trend line. The sample window is 1% of the number of sequences considered. To further illustrate the spreading of relatively cyclic timetables resulting from the simulations, all sequences with $\overline{C_{ij}} \leq 10\%$ are included as well. This corresponds approximately to 45% of the total number of sequences. Using 60 seconds as minimum resolution, this number drops to 40%.
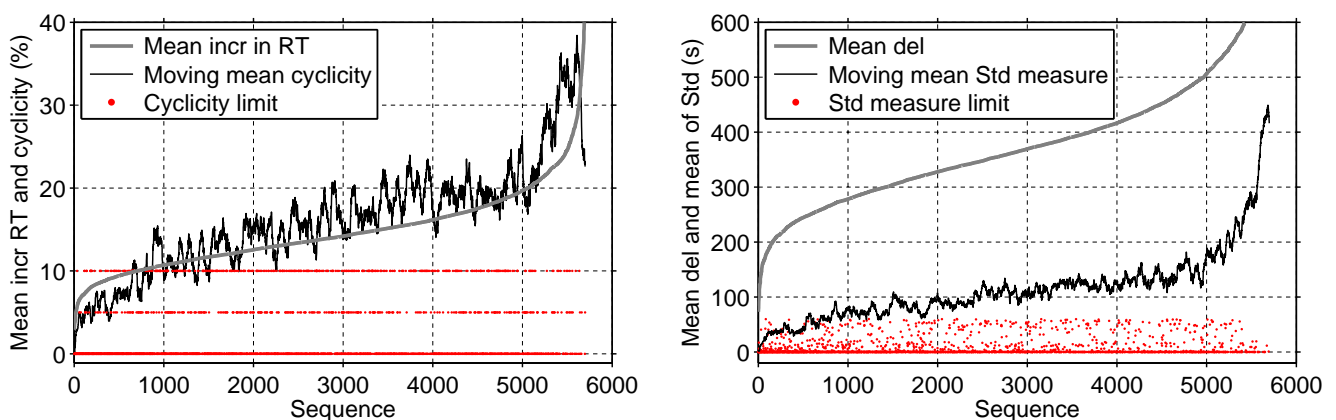

Figure 6: Measures of total cyclicity, mean for all groups, based on equation 2 and standard deviation of scheduled delay. Moving mean indicates trend when sorted with respect to increasing mean delay for all groups.

Representing cyclicity with the mean of group standard deviations shows similar behaviour as the already discussed cyclicity measure. Sequences with a maximum value of 60 seconds are included in figure 6, around 40% of the total. Lowering the maximum deviation value to 30 seconds gives 36%. Same conditions as described previously apply to the moving mean. This representation of the spread in run time or scheduled delay can, in combination with other measures, be applied in choosing sequences for further studies.

More weight can be given to some train groups with respect to other groups in the simulations, thus producing other timetable solutions with unchanged start sequences. How much weight is given to priority with respect to total expected delay when comparing two dispatching solutions can be varied in RailSys. This is done by considering the expected added delay difference for two trains and comparing this to a threshold percentage value which can be varied. A higher threshold percentage means that more weight is given to trains with higher priority than to total added delay.

Figure 7 shows the change in mean percentage of increased run time for a train group when more dispatching priority is given to the other group. To a varying degree this means that the group with lower priority is more likely to face longer waiting times in meets and overtakings, also increased number of these events in some situations. On the contrary, trains getting increased priority are expected to experience shorter waiting times and less obstruction from trains with lower priority. Comparing sequence by sequence reveals that high speed trains, as anticipated in most cases, benefit from this (figure 7 right) and get lower increased run times.
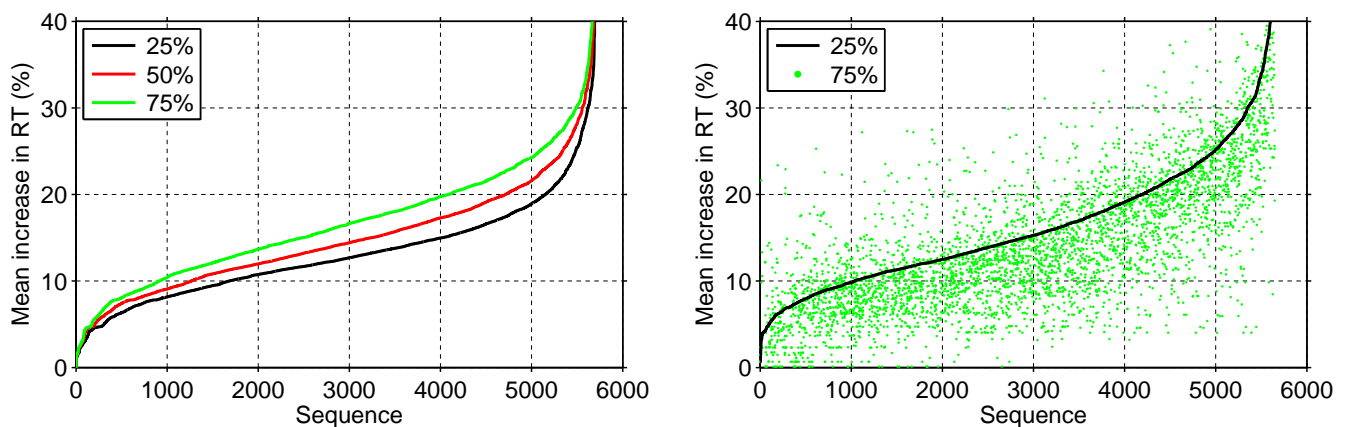

Figure 7: Effects on mean percentage of increased run time when dispatching priority is raised for high speed trains (HP). Left figure shows results for regional trains (RP) sorted individually for each case. Right figure shows sequence-by-sequence results for high speed trains.

Changing a train meet to an adjacent station can give a high impact on the waiting time and consequently on the total run time. In this case, high speed trains on average experience shorter waiting times due to their higher weight in dispatching decisions. It should be emphasized that in figure 7 both directions are combined for respective train group and due to the interactions effects characteristic for single tracks, decisions that affect one train negatively at some location can later lead to a better solution (or vice versa) for a train running in the opposite direction belonging to the same group. These interactions also explain why not all sequences result in equal or lower values in the right figure.

5.2 Case with passenger and freight trains

Based on the results shown in figure 5, 30 sequences giving the lowest percentage values of increased run time are selected as input for a second step including a third train group. In other words, requested start start times included in the first step are fixed. The start sequences form a new set of preconditions and freight trains (FR), with characteristics described in figure 4 and table 2, are added using the combinatorial approach explained in section 4. In addition a limit condition is applied, freight train time slots are varied in separate hours with respect to direction, hence all possible combinations are not captured. Minimum initiation headway is nine minutes for trains succeeding a freight train and six minutes in other cases, setting limitations for available time slots.

Generating sequences with these conditions give 4528 combinations in total and after simulation 2610 feasible solutions are used for evaluation. The difference to the number of input start sequences is discussed in section 6. The mean percentage values for increased run time are presented in figure 8, all three evaluation groups are sorted individually from lowest to highest value giving curves similar to figure 5.

Sorting regional and freight train 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.
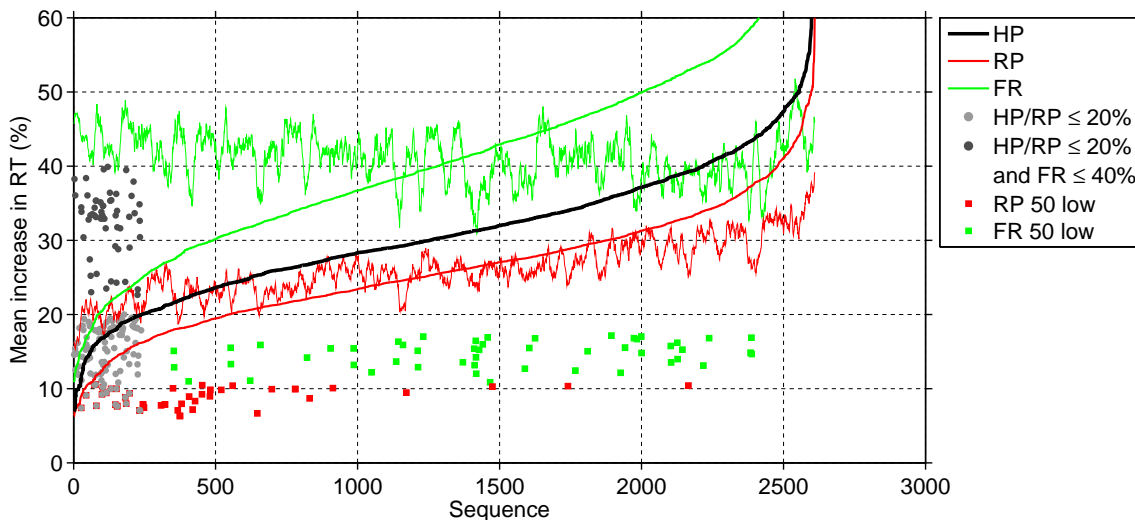

Figure 8: Two track station layout. Mean percentage of increased run time, sorted individually for each group and according to high speed trains (HP) as a moving mean with sample window 1%. Timetables not exceeding specified limits (%) are indicated together with timetables giving the 50 lowest group values.

Assuming that some of these sequences become input to simulations with perturbations in order to evaluate the effect of delay propagation (section 1.1 and 1.2), threshold values can be used in finding preferred timetable solutions. Figure 8 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.

Figure 8 also indicates that freight trains on average experience longer waiting times than the higher prioritized passenger trains as expected. Having no pre scheduled stops means that all stops taking place in the simulations always contribute to an accumulated delay. Passenger trains can simultaneously have scheduled stops and meets or overtakings, thus some of the waiting time coincides with scheduled dwell time. Choosing a sequence and plotting actual arrival and departure times in a time-distance diagram (graphical timetable) gives a good idea of the traffic situation. The best timetable considering a mean value for both high speed and regional trains is shown in figure 9. High speed trains have a scheduled stop at station F and regional trains at stations D, F and I. Since the 30 best sequences were kept from the previous run, the benefits of combining meets and scheduled stops are clearly seen in this example.
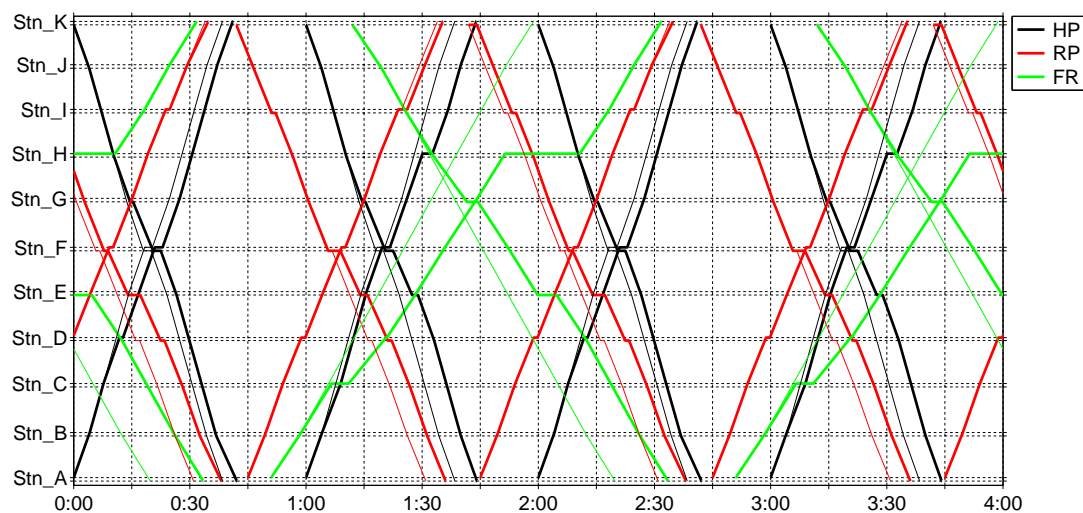

Figure 9: Simulation results for one start sequence with three train groups and two track station layout. Bold lines show simulated (conflict managed) train paths and thin lines pre-scheduled paths.

5.3   Case with extended station capacity

In order to achieve more flexible conditions for meets and overtaking situations a third track (a second siding) is added to all intermediate stations. This will in theory allow three train meets or combination of meets and overtakings at the same time. The added station tracks are equivalent with the existing sidings considering track lengths and diverging speeds in switches. Comparing results with the same sequence pre-conditions but with different station capacity, shows that the number of feasible solutions increases with almost 70% (from 2610 to 4417). For the three track case, figure 10 illustrates the same measures as explained in the previous section.

Comparing with figure 8, the difference between the sorted passenger and freight train curves is greater. One explanation is that while the number of solutions increases, the additional solutions became deadlocked in the two track case due to conflict situations involving several trains. Adding

a third station track gives more degrees of freedom meaning that lower prioritized trains will face waiting time allowing for meets and overtakings between higher prioritized trains. The dispatching conditions for passenger trains improve and lead to lower scheduled delay or total increased run time in sequences giving solutions in both station layout cases and also in the additional ones.
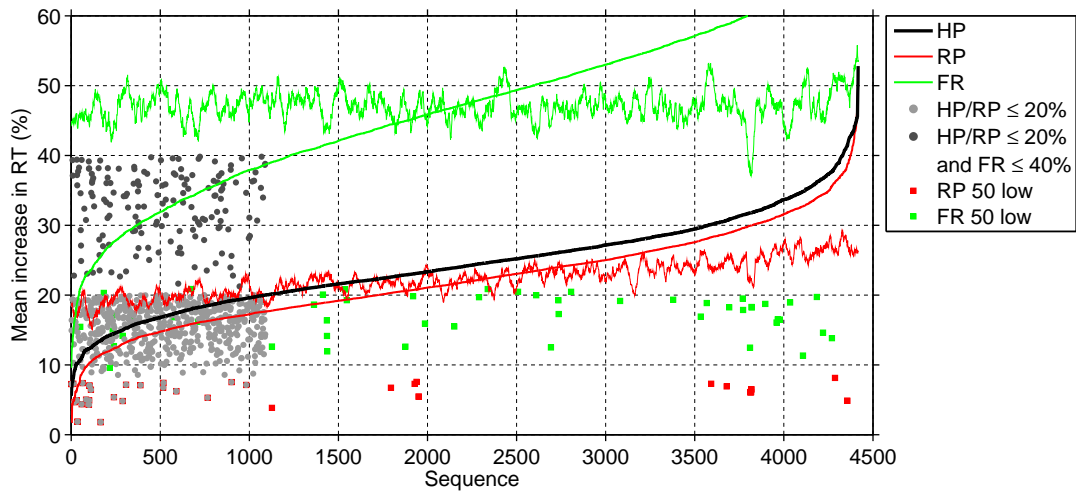

Figure 10: Three track station layout. Mean percentage of increased run time, sorted individually for each group with respect to high speed trains (HP) as a moving mean with sample window 1%. Timetables not exceeding specified limits (%) are indicated together with timetables giving the 50 lowest group values.

Applying equal conditions as in the previous example regarding maximum allowed scheduled run time increase result in 1090 timetables satisfying HP $\leq$ 20%. Considering these timetables, 617 also meets the requirement RP $\leq$ 20% and 167 if the requirement FR $\leq$ 40% is added. The interrelationship between the number of timetables considering high speed and regional trains is approximately same. Relatively speaking fewer solutions fulfilling the limit for freight trains come out in the three track case.

Figures 8 and 10 show that the number of solutions for freight trains is nearly equal up to 40%, i.e. the three track case contributes in creating solutions above this limit. Figure 11 shows more clearly that the situation mostly improves for passenger trains, whereas freight trains get increased run times in many timetables given the same requested start times.
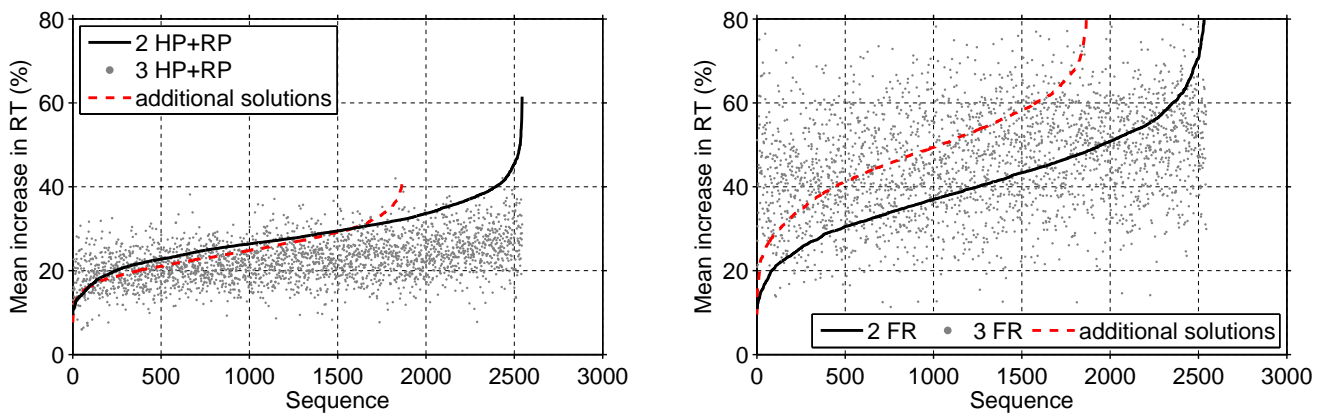

Figure 11: Cases with two and three track stations, left figure shows passenger trains combined and right figure shows freight trains. Dots show the spreading of percentage values if cases are compared sequence by sequence, with respect to the solid line. Additional solutions obtained in three track case shown by dashed lines.

Considering the additional solutions obtained in the three track case they contribute in improving the overall sorted curves for passenger trains (figure 10), meaning that the number of timetables meeting a certain percentage limit increases. However, as previously mentioned, for freight trains the spreading of sequence-by-sequence and additional values give no significant difference up to 40% increased run time. The best timetable considering a mean value for both high speed and regional trains is shown in figure 12. In this example the third station track is utilized at station D and H, also it does not coincide with the same sequence as in figure 9. Visually comparing the difference between pre-scheduled and simulated freight train paths in these two timetables gives that freight trains have more increased run time in the three track case.
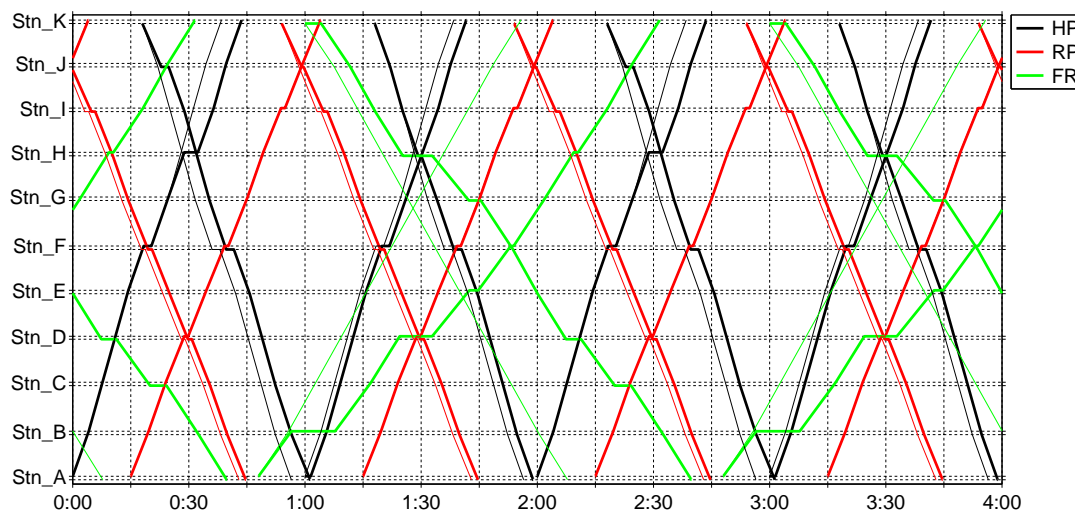

Figure 12: Simulation results for one start sequence with three train groups and three track station layout. Bold lines show simulated (conflict managed) train paths and thin lines pre-scheduled paths.

Attempts with freight trains running one time every hour per direction using the initial two track layout gave relatively few feasible timetables compared with the number of sequences. It should be emphasized that running three trains per hour and direction (six in total) with this speed heterogeneity implies a high track and station occupation on the modelled infrastructure. However, increasing the infrastructural flexibility could give possibilities of running more trains within the same time. Using same sequences as in the two track case, rendered significantly higher number of feasible timetables.

6   Conclusions

Evaluating a set of timetables with similar properties regarding the number of trains running, train characteristics and so on can give valuable insight in itself and also be useful in simulation studies including stochastic and systematic perturbations reflecting real conditions. A timetable may appear attractive beforehand and perform poorly when executed in reality. Simulating with applied perturbations could give a forecast of this and, if done for several alternatives, give guidance into which timetables are preferred.

This approach with start sequences can be used both on single and double tracks. If the directions on a double track can be assumed to be almost independent, the combinatorial problem becomes easier to handle. However, if many interactions exist with trains crossing tracks these will influence the overall timetable to some degree and measures may be necessary in the sequence generation setup to capture this effect. An obvious limitation is the number of sequences that are reasonable to simulate with respect to computation time. Therefore effort is required in the setup to reduce

the problem and still cover a wide range of alternatives. Since the results are not known until all sequences are simulated, cases which lead to more or less unrealistic solutions consume computation time as well.

One way of partially overcoming the problem with limitations in the number of manageable sequences is to repeat the process after a first simulation and use some solutions as input to a second run with sequence generation. More trains can be added, existing sequences varied with finer time resolution and so on with the objective of finding more or better solutions for stochastic simulations and other studies. However, this is not a method giving all possible solutions (timetables) and it is therefore difficult to know if and how many feasible alternatives exist not captured by this method.

Another possibility could be to generate all possible sequences given some pre-conditions and randomly choose a manageable number of sequences prior to simulation. This could result in the same typical curve characteristics for sorted values as seen in several figures in this paper, meaning a low number of timetables with low and high values compared to the total. Assuming that timetables found in this way will be used in some continuation or other study, only a limited number of these will be of interest. This means that a majority of the timetables are of no interest due to values of increased run time beyond what is acceptable.

There are also limitations in the dispatching capability of the simulation software; the routing function is heuristics-based and not optimisation-based. In particular, single track lines with dense heterogeneous traffic are complicated from a dispatching point of view. Once an expected or current conflict is detected the software can choose from different dispatching methods depending on the situation. However, sometimes a conflict situation cannot be resolved by the routing function, thus leading to deadlocks. This is due to the limitations described and the fact that none of the trains running in these simulations are conflict managed, since the main point is to use RailSys for generating many different timetables. These will then by definition be conflict free, by that not claiming they are automatically realistic.

The run times, including scheduled stops, used in this study have almost no allowance. In order to obtain feasible timetables with built in slacks (margins), these can be introduced for the template trains prior to simulation. This gives more resilient or robust timetables in simulations with stochastic perturbations. The slacks can then, to a varying degree, be used for reducing the effects of perturbations and knock-on delays. Considering the relatively low number of cyclic passenger train timetables with respect to limit values in cases with freight trains, more detailed analysis of this is needed. For example, other measures describing cyclicity can be evaluated. In the case where freight trains are running with half the frequency of passenger trains, the mean standard deviation goes up since the meet and overtaking situations vary for two consecutive passenger trains belonging to the same group.

Handling cases where all trains in one direction do not start at the same station is possible. The headway criteria become more imprecise since assumptions are needed on expected run times between specific stations. In this study delay starts ticking from the sequence initiation time even if a train cannot depart directly due to other trains. Taking the actual departure time as a reference for delay calculations can give more accuracy. However, considering the combinatorial properties this situation may well be captured in another sequence depending on the time parameters in use. It is also possible to use some sort of randomization in the initiation process, hence not using regular interval departures. Mainly this could add more realism in freight operations.

The results in this paper show that a selection of timetables can be obtained by combining systematic train initiation times as input to RailSys and taking advantage of the software's ability to handle conflicts between trains. Further investigations are needed in order to assess this method and draw more general conclusions. As previously mentioned, the main motive for this work is

to evaluate timetables by simulating several comparable alternatives including perturbations and rank them. Naturally this requires some degree of accuracy in modelling realistic perturbations. Simulating only one timetable can of course give valuable insight and answer some questions. The credibility of conclusions and recommendations from a simulation study will benefit if it involves several timetables.

7   References

Dingler, M.; Lai, Y.-C.; Barkan, C.: Impact of train type herogeneity on single-track railway capacity. In Journal of the Transportation Research Board, March 2009.

Koutsopoulos, H.; Wang, Z.: Simulation of urban rail operations: Application framework. Proceedings of the Transportation Research Board, 2006.

Koutsopoulos, H.; Wang, Z.: Calibration of urban rail simulation models: A methodology using SPSA algorithm. In Proceedings of the 2011 Winter Simulation Conference, WSC, pp 3699–3709, USA, 2011.

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

Lindfeldt, O.: Analysis of capacity on single-track railway lines. In Proceedings of 4th International Seminar on Railway Operations Modelling and Analysis, Rome, Italy, 2011.

Mattsson, L.-G.: Critical Infrastructure: Reliability and Vulnerability. Springer-Verlag, ch. 7, pp 129–150, 2007.

Olsson, N.; Haugland, H.: Influencing factors on train punctuality results from some Norwegian studies. In Transport policy 11, pp 387–397, 2004.

Radtke, A.; Hauptmann D.: Automated planning of timetables in large railway networks using microscopic data basis and railway simulation techniques. Computers in Railways IX. WIT Press, 2004.

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

Siefer, T.: Simulation. In: Hansen, I. A.; Pachl, J. (editors): Railway timetable and traffic, pp 155–169. Eurail press, 2008.

Sipilä, H.: Calibration of simulation model on the Southern Main Line In Sweden. Railway Engineering – 2011, 11th International Conference, London, UK, 2011.

Sogin, S.; Barkan, C.; Saat, M.: Simulating the effects of higher speed passenger trains in single track freight networks. In Proceedings of the 2011 Winter Simulation Conference, WSC, pp 3679–3687, USA, 2011.