Journal of Rail Transport Planning & Management 3 (2013) 111–123

Contents lists available at ScienceDirect

Journal of Rail Transport Planning & Management

journal homepage: www.elsevier.com/locate/jrtpm

Maintaining tracks and traffic flow at the same time

Malin Forsgren *, Martin Aronsson, Sara Gestrelius

SICS Swedish ICT, Box 1263, SE-164 29 Kista, Sweden

ARTICLE INFO

Article history:
Received 10 September 2013
Accepted 3 November 2013
Available online 1 February 2014

Keywords:
Timetabling
Conflict minimization
Track possession planning

ABSTRACT

In an ideal world, all railway tracks would be available to trains at all times. In reality, track sections need to be closed every now and again for track maintenance and upgrades in order to ensure a satisfactory level of safety and comfort. In this paper, we present a MIP model that optimizes a production plan with regard to both trains and preventive maintenance. The planned maintenance activities may not be canceled, but may be moved in time within pre-defined time windows. Trains may be moved in time, redirected to other parts of the geography, or even canceled. The goal for the optimization is to find the best possible traffic flow given a fixed set of planned maintenance activities. In addition to presenting the model, we discuss the current maintenance planning process in Sweden, and exemplify the usefulness of our model in practice by applying it to two typical scenarios.

1. Introduction

The basic structure of a yearly train timetable assumes that the infrastructure is intact and in good shape. In reality, track sections need to be closed every now and again for track maintenance and upgrades in order to ensure a satisfactory level of safety, comfort and future availability. Choosing the most suitable times for closing tracks for maintenance, so called *track possessions*, is a tough challenge that all infrastructure managers face today.

Corrective maintenance schedules itself: when an acute infrastructure problem arises, the problem simply has to be dealt with as soon as possible. Preventive maintenance, on the other hand, can be planned long in advance. For track sections with dense traffic, it might not be possible to schedule all required track possessions to times when the tracks are not needed by trains. As a result, track possessions and train paths have to fight for the same capacity.

While there are many problems that need to be solved in relation to planning for preventive maintenance, see Section 2, the focus of this paper is how to schedule track possessions in a given railway timetable. We have developed a model that schedules track possessions alongside trains in such a way that all the maintenance activities can be performed, while as much as possible of the traffic flow of the original timetable is maintained.

Our model assumes that there always exists a plan that best fulfills the goals for the given traffic, and that these traffic goals are represented fairly by the yearly plan, rolled out for a specific time period, and updated with regard to which trains have been added or canceled since it was published.

The best production plan for the trains is the one without any kind of disturbances. For this reason, we keep the number of track possessions fixed in the model: As soon as capacity becomes scarce, freeing up capacity by canceling track possessions, or postponing possessions beyond the time horizon of the current production planning period, would otherwise always give a better plan for the trains.

An overview of related work can be found in Section 2. The current planning practices in Sweden are briefly discussed in Section 3. The basic timetabling model is defined in Section 4, and Section 5 that follows describes the concepts needed for including track possessions in the model. The complete model formulation is included as Appendix A. Sections 6–8 discuss the implementation of the model, the objectives of the optimization, and input data considerations. Last we present practical results for typical scenarios in Section 9 and end with a summary and suggestions for future work in Section 10.

2. Related work

Compared with the number of papers published on the train timetabling problem, there are very few academic papers published on planning with both trains and track possessions (see the literature overview in Budai-Balke, 2009). In addition, most of the papers that do consider both trains and track possessions focus on scheduling one of them while the other is viewed more or less as a side constraint.

Our paper describes a model that is capable of dealing with a realistic scenario close in time to the real-time operations, where

* Corresponding author. Tel.: +46 (0) 8 6331594.
E-mail addresses: malin@sics.se (M. Forsgren), martin@sics.se (M. Aronsson), sarag@sics.se (S. Gestrelius).

2210-9706/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jrtpm.2013.11.001

both trains and track possessions obviously need to be considered. While our model does not give trains and track possessions completely equal treatment, it nevertheless schedules them simultaneously. The only approach so far that we are aware of that does schedule maintenance and trains simultaneously is the Australian proposal to the problem of scheduling long-haul single-track networks (Albrecht et al., 2013; Pudney and Wardrop, 2004). This method is however not directly applicable to the Swedish (or European) situation, as the Australian network is mostly used for freight trains that do not have rigid timetables to adhere to.

There are several aspects to scheduling maintenance. One interesting aspect is how to be able to best predict the need for preventive maintenance. Research in this area focuses on the strategic, long-term perspective (Putallaz and Rivier, 2003), or the yearly maintenance (Cheung et al., 1999). The tear and deterioration of infrastructure components in the railway domain has received a lot of attention (see e.g. Andrade and Teixeira, 2012; Larsson, 2004; Simson et al., 2000; Plu et al., 2009), and the general topic of calculating the maintenance frequency when a model for the respective wear is known has been thoroughly studied. For an overview of the latter, see (Jardine et al., 2006).

Another aspect concerns the maintenance activities as such, e.g. focusing on how to perform the activities required to take care of an underlying maintenance need in the most efficient way. Efficiency in this context can mean two things: actual cost (in money), and how much the traffic needs to be disturbed. If traffic disturbance can be expressed as a cost, both these aspects can be considered simultaneously (see Budai et al., 2004).

One of the most common approaches to reduce the cost of maintenance is to find strategies to lump different activities together in maintenance packages (Budai et al., 2004; Vatn, 2008; Peng et al., 2011; Peng and Ouyang, 2012). Assuming that the cost of the performed activities does not vary depending on when they are carried out, the cost can be reduced by minimizing the overhead in terms of paying salaries and moving crews and equipment in the geography. Indirect costs due to canceled or redirected trains are not considered in these models.

Research that explicitly focus on minimizing the disturbance to the traffic is rare, but (Lake and Ferreira, 2002) falls into this category. For research that does not explicitly minimize the traffic disturbance, but still takes it into consideration by adding certain side constraints, we would like to mention (Peng et al., 2011). In (Budai et al., 2004), the authors are aware that the trains should not be disturbed, but they argue that what is best from the possession planners' perspective is also good enough for the traffic situation: To group the work shifts and keep them together instead of splitting them is the cheapest way to get lots done since it lessens the need for set-up times, and in a general sense it is advantageous also from the point of view of the traffic since it minimizes the total time the tracks are unavailable.

Among the methods used to solve the actual scheduling problem, we have found examples of MIP models (Fischetti et al., 2007; Lake and Ferreira, 2002), Constraint Programming (Cheung et al., 1999) and Genetic Algorithms (Budai-Balke et al., 2009; Higgins and Kozan, 1997). Heuristics such as Tabu Search (Budai-Balke et al., 2009; Lake and Ferreira, 2002; Pacciarelli and Pranzo, 2001), Local Search (Higgins and Kozan, 1997; Lake and Ferreira, 2002), and Simulated Annealing (Lake and Ferreira, 2002) are also used.

3. Current planning practices in Sweden

Deciding what changes should be made to an existing train plan to accommodate for more track possessions is a complicated process in which many different factors have to be considered. In order to make room for track possessions, some trains will have to be moved in time, get longer running times, be canceled, be assigned to other routes in the network, or suffer a combination of these measures.

In Sweden, infrastructure maintenance is outsourced. In short, this means that, for certain kinds of preventive maintenance activities, the infrastructure manager (IM) does not decide how or when they should be carried out. For other types of jobs, the IM will make a rough plan and suggest both method, and time and date for the job, although the IM will not be sure that there will be any maintenance entrepreneur capable and available to carry out the job as planned.

The major task of a track possession planner at the IM is to negotiate with RUs (Railway Undertakings) and entrepreneurs in order to find suitable times for the maintenance jobs. He or she can often guess what alternatives the RU will consider for their trains, should they be affected by track possessions. Constrained by the proposed method, the time windows for when maintenance crews are available, and all other relevant constraints, the planner will decide a placement in time for the track possession that, to the best of his/her knowledge, and given that he/she can predict accurately what the RUs want to do with the affected trains, causes the least disturbance to the traffic flow.

Once it has been established that a track possession will interfere with a given train path in the timetable, it is normally entirely up to the RU to apply for an alternative train path, or to start planning for canceling the train on the affected dates. If the RU decides to apply for a new train path, the RU is supposed to apply for a train path that fits into the current train plan, but naturally lacks the full picture since the RU is not aware of what alternative train paths other RUs are simultaneously planning to apply for. The IM might therefore have to modify the RU's application in order to be able to schedule it on available capacity without interfering with existing train paths on the tracks in question.

The applications for alternative train paths are usually treated by the IM on a first-come, first-served (FCFS) basis, although the IM might make an exception to this principle when it comes to track possessions in areas with dense traffic where capacity is particularly scarce. In such a situation, the FCFS principle risks wasting too much capacity compared with considering many applications for alternative train paths at the same time; the IM might instead set a deadline for the applications from all affected RUs. After the deadline, a more careful planning process takes place in order to create a new production plan with minimum disturbance to the traffic as a whole.

When the IM has constructed an alternative train path for the RU, the RU has to decide whether it is feasible to use the alternative train path or not. The answer to that question depends on the properties of the new train path.

The inherent delay for the train brought upon the RU by the mere fact that the alternative route is longer, or at least slower for the train, does of course not come as a surprise to the RU, nor does the alternative route itself. After all, the RU stipulated these properties in their application. Less predictable delay caused by interference with other trains might however cause the RU to choose to cancel the train rather than to redirect it. Also, it is possible that the IM has had to change the alternative train path applied for significantly in other ways than just delaying it. E.g., the RU might want the alternative train path to have a particular departure or arrival time at some specific location, and if the IM cannot meet this request, the alternative train path may no longer be commercially feasible to the RU.

4. The basic timetable model

This section describes the basics of the MIP model we use to handle the scheduling of trains. Section 5 extends the model to

include the notion of redirecting trains to accommodate for track possessions. The complete mathematical formulation of the full model can be found in the Appendix A.

Our research is funded by the Swedish IM Trafikverket, the Swedish Transport Administration (previously Banverket, the Swedish Rail Administration), and our aim is to develop methods and tools that they can use in their organization. Trafikverket use the timetabling tool TrainPlan (offered by Trapeze, previously Funkwerk IT) to store, maintain and display relevant data, e.g. to visualize the train plan as time–distance graphs. TrainPlan does not have any optimization functionality. The similarities between our model and the underlying model of TrainPlan are intentional, as we naturally want to work with the same data and produce optimized plans that are easy to understand for people who are familiar with TrainPlan.

4.1. Locations and links

The railway network is modeled as *locations* and *links* and can be likened with a bidirected, connected graph with nodes (locations) and one or many arcs (links) between neighboring nodes.

Locations represent stations and other places in the network (e.g. railroad switches and sidings) for which we want to be able to assign departure and arrival times for trains. Every link represents a unique track section between two locations; the existence of double-track or multi-track between two locations is represented by one link per physical track.

4.2. Routes

We formally define a *route r* in the network as an ordered set of consecutive links whose physical correspondence in the real world can be traversed (in that order) by a train. The locations along the route, including the first and last ones, are called the *route locations*.

A *valid route* is a route in the network that a train would normally be able to traverse. The order in which the two locations of a link are visited in a route decides which location is called the start and end location of that link, when associated with that particular route.

Analogously, the start location of the first link of a route is called the *start location of the route* and the end location of the last link is called the *end location of the route*.

4.3. Trains and track possessions

We define a train *i* as a valid route *r*, a calendar, a train type, and a set of location activities for the route locations. The calendar tells on what dates the train will depart. A *location activity* describes whether a train has a planned activity at a location *g* or not, and if it does, specifies the minimum duration w_g^i of the associated stop.

Trains are represented by chains of alternating location activities and link traversals. The pair of equations (1) below define the relationship between the arrival at *C*, the departures on links R_3 and R_4, the trip time on link R_3, and the minimum dwell time w_C^i at location *C* of train *i* in Fig. 1.

$$\begin{aligned} d_{R_3}^i + t_{R_3}^i &= a_C^i \\ a_C^i + w_C^i &= d_{R_4}^i \end{aligned} \tag{1}$$

The minimum trip time on a link for the train depends on the train type and the location activities on the start and end locations of the link, see Section 4.4. Given the location activities of the train,

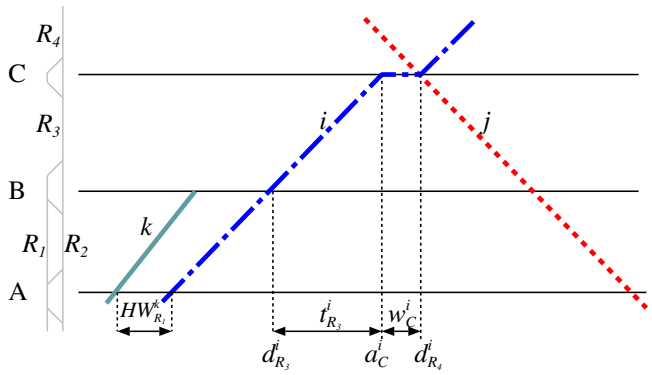

Fig. 1. Three trains *i*, *j* and *k* scheduled in a network with locations *A*, *B* and *C* and links R_1, R_2, R_3 and R_4, with physical layout according to the sketch to the left of the time–distance diagram.

its *nominal running time* is the sum of the minimum trip times on the links, and the minimum dwell time durations at the associated route locations where the train has planned stops.

A track possession is defined as an activity on one or more links that has a specified, fixed duration. Every track possession has one or many suggested dates (calendars) and start times. A track possession "takes possession" of all the affected links at the same time, and gives them back all at the same time when the possession ends.

4.4. Link trip times

The signaling system in Sweden supports full capacity bidirectional operation on both tracks on double-track lines. For a certain train type, there are four possible minimum trip times for every link and direction, given by a database of running times. They represent the minimum time required for a train of that particular train type to traverse the link when it is either entering and exiting the link at full speed (*FF*), starting from stop and exiting at full speed (*SF*), entering at full speed and stopping at the end location of the link (*FS*), or stopping at both link locations (*SS*).

Since a train belongs to exactly one train type, we can denote the minimum trip time $T_{\lambda,X}^i$, where *X* denotes the stopping behavior *FF*, *SF*, *FS* or *SS* of train *i* on link λ as described above. Note that the travel direction of a train is defined since the train is associated with a specific route.

We let $g_1(i,\lambda)$ and $g_2(i,\lambda)$ denote the start and end locations for a link $g_2(i,\lambda)$ when traversed by train i. Each train i will have its trip time t_λ^i modeled in one of the following four possible ways for every link it traverses, depending on whether the stopping behavior of the train is predefined or not, at one, both or none of the link locations.

1. The stopping behavior is predefined at both the start and end locations. The minimum trip time is simply fetched from the database, and the trip time is expressed as

$$t_\lambda^i - s_\lambda^i = T_{\lambda,X}^i$$

where s_λ^i is a slack variable.

2. The stopping behavior is predefined at the departure location but not at the arrival location. In this case, we get one of the two following equations, the choice of which is governed by whether the train has made a stop or not at the departure location:

$$t_\lambda^i - s_\lambda^i + \left(T_{\lambda,SF}^i - T_{\lambda,SS}^i\right) z_{g_2(i,\lambda)}^i = T_{\lambda,SF}^i$$

or

$$t_\lambda^i - s_\lambda^i + \left(T_{\lambda,FF}^i - T_{\lambda,FS}^i\right) z_{g_2(i,\lambda)}^i = T_{\lambda,FF}^i$$

where z_g^i is a binary variable that will be assigned the value 1 if train i stops at location g. To ensure that z_g^i is 1 when the dwell time w_g^i is non-zero, we require

$$w_g^i - Mz_g^i \leq 0$$

to hold per train i and location g, where M is a constant bigger than w_g^i.

3. The stopping behavior is predefined at the arrival location but not at the departure location. This is analogous to case 2 above, with the roles reversed. Thus, either

$$t_\lambda^i - s_\lambda^i + \left(T_{\lambda,FS}^i - T_{\lambda,SS}^i\right) z_{g_1(i,\lambda)}^i = T_{\lambda,FS}^i$$

or

$$t_\lambda^i - s_\lambda^i + \left(T_{\lambda,FF}^i - T_{\lambda,SF}^i\right) z_{g_1(i,\lambda)}^i = T_{\lambda,FF}^i$$

will be used.

4. The stopping behaviors at both the start and end locations are unknown. We introduce four new variables $f_{\lambda,X}^i$, and require

$$\begin{aligned}
f_{\lambda,FF}^i + f_{\lambda,SF}^i + f_{\lambda,FS}^i + f_{\lambda,SS}^i &= 1\\
f_{\lambda,FF}^i + z_{g_1(i,\lambda)}^i + z_{g_2(i,\lambda)}^i &\geq 1\\
f_{\lambda,SF}^i - z_{g_1(i,\lambda)}^i + z_{g_2(i,\lambda)}^i &\geq 0\\
f_{\lambda,FS}^i + z_{g_1(i,\lambda)}^i - z_{g_2(i,\lambda)}^i &\geq 0\\
z_{g_1(i,\lambda)}^i + z_{g_2(i,\lambda)}^i - f_{\lambda,SS}^i &\leq 1\\
s_\lambda^i - t_\lambda^i + T_{\lambda,FF}^i f_{\lambda,FF}^i + T_{\lambda,SF}^i f_{\lambda,SF}^i + T_{\lambda,FS}^i f_{\lambda,FS}^i + T_{\lambda,SS}^i f_{\lambda,SS}^i &= 0
\end{aligned}$$

to hold. Note that $f_{\lambda,X}^i$ need not be binary declared.

4.5. Train paths

The commonly agreed upon definition of the term *train path* reads "the infrastructure capacity needed to run a train between two places over a given time-period" (The European Parliament and the Council, 2001). With the definitions used in this paper, a train path is associated with a valid route and a set of specific departure and arrival times at all route locations.

4.6. Trains and their timetables

Our definition of a train does not involve specific arrival and departure times. Instead, constructing a timetable means assigning non-conflicting train paths to trains for the dates in the trains' calendars. Consequently, some properties, such as e.g. the arrival time of a train, only exist in the context of a specific timetable.

The *scheduled departure/arrival time* of a train at a location is the departure/arrival time assigned to the train at that location in a given timetable. The *scheduled running time* of a train is the difference between the scheduled arrival time of the train at its end location and the scheduled departure time at its start location.

The scheduled running time has three components:

1. The nominal running time
2. Time supplements
3. Pathing time

A *time supplement* is additional time added to compensate for small, everyday variations in j performance, to make the train plan less sensitive to varying weather conditions, different driver behaviors, etc. *Pathing time* is the time added to trains at locations (or on links) during the timetabling process as a means of resolving resource conflicts, and is thus the time the train is scheduled to spend waiting for its turn to get access to the tracks.

Time supplements are part of the input data and considered fixed during optimization, whereas pathing time may be both subtracted and added. By specifying at which locations pathing time may be added to which train, trains can be prevented from being scheduled for pure technical stops where this would not be suitable.

We have approximated the blocking time theory to facilitate computing resource conflicts without the need for very detailed input data. A link on a single-track line is thus assumed to be blocked for train j by the preceeding train i for the duration of the trip time of train i and an extra amount of time that depends on the relative directions of trains i and j, the type of locations (for trains with opposing directions), and the respective location activities of the trains. For links on a double-track line, we assume that the signals are close enough to enable a headway approximation for the separation of trains moving in the same direction. If trains move in opposing directions on a link that is part of a double-track line (remember that every track is a separate link), they naturally occupy the link for the duration of their trips in the same manner as if the link constituted a single-track line.

Resource conflicts on links are regulated with a big-M formulation, using the binary variable x_λ^{ij} to ensure that one of the trains i and j starts traversing the link λ before the other, and that they are separated adequately. For trains with opposing directions on the same link belonging to a double-track line, and generally on single-track lines regardless of relative direction, the basic inequalities that must hold are

$$\begin{aligned}
d_\lambda^j - d_\lambda^i - t_\lambda^i + M(1 - x_\lambda^{ij}) &\geq 0\\
d_\lambda^i - d_\lambda^j - t_\lambda^j + Mx_\lambda^{ij} &\geq 0
\end{aligned} \tag{2}$$

For trains traveling in the same direction on the same link on a double-track line, a separation of the departures has to be maintained at the start location of the link, and the two trains have to be separated also at the end location. The headway HW_λ^i (see Fig. 1), is potentially train specific. Thus,

$$\begin{aligned}
d_\lambda^j - d_\lambda^i - HW_\lambda^i + M(1 - x_\lambda^{ij}) &\geq 0\\
d_\lambda^i - d_\lambda^j - HW_\lambda^j + Mx_\lambda^{ij} &\geq 0\\
a_{g_2(i,\lambda)}^j - a_{g_2(i,\lambda)}^i - HW_\lambda^i + M\left(1 - x_\lambda^{ij}\right) &\geq 0\\
a_{g_2(i,\lambda)}^i - a_{g_2(i,\lambda)}^j - HW_\lambda^j + Mx_\lambda^{ij} &\geq 0
\end{aligned} \tag{3}$$

must all hold.

The capacity of each location is approximated by an integer dictating how many trains that the location can host at the same time. To ensure that the capacity at locations is respected, we use a model called the min conflicting sub-clique model. A detailed account of this model can be found in a previously published paper of ours (Aronsson et al., 2009).

Mathematically, a timetable is any assignment of arrival and departure times that respects the constraints for the trains, the links, and the locations given or referred to in this section.

5. The extended model

Section 4 describes the basic MIP model used for timetabling. This section adds definitions that are needed for enabling the rescheduling, and possibly redirection, of trains in the event of track possessions.

In this section, we assume that there is a published train plan, and that the goal is to make temporary modifications to it for selected dates in order to make room for one or more track possessions. The new plan found, for the selected dates, when solving the MIP, will be referred to as a *suggested timetable solution* to

emphasize that the plan has not officially replaced the published train plan yet.

We let an *original train* be a train in the published train plan. The *original train path* of the original train is the train path that the original train was assigned to in the published train plan. The departure and arrival times of a specific original train in the published train plan will accordingly be referred to as its *original arrival and departure times*.

5.1. Alternative routes

We call the routes r_a and r_b *alternative routes* if the following assumptions hold.

- r_a and r_b are valid routes
- r_a and r_b have the same start and end locations
- At least one of the links in the respective routes differs

See Fig. 2 for two examples of alternative routes.

5.2. Alternative train paths

We call p_{r_a} and p_{r_b} alternative train paths if

- r_a and r_a are both valid routes with the same start and end locations
- p_{r_a} and p_{r_b} have different arrival times or different departure times at one or more locations

Note that the routes r_a and r_b can be either identical or alternative routes. The key is that they do not have the same arrival and departure times at all locations. This means that any change to an original train's departure or arrival times after the train plan has been published formally requires a new, alternative train path for the train.

5.3. Train versions

Remember that a train is not associated with absolute departure or arrival times until it has been assigned to a train path. Let i_0 denote the *original version* of the original train, meaning that it uses the same route as the original train, has the same location activities, and the same nominal running time.

An *alternative train* is a train that could potentially replace an original train for the duration of a track possession, but that uses a different route compared with the route of i. Note that this means that an alternative train can (but does not need to) pass the same locations as i, and it can have a different nominal running time from that of the original train's.

Finally, we introduce the concept of the canceled train. A *canceled train* can be viewed as an imaginary train that could potentially replace a specific original train i for the duration of a track possession. However, the canceled train does not use any route.

The alternative trains of i, the canceled train, and the original version of i are collectedly referred to as the *train versions* of i. A suggested timetable solution will consist of precisely one version of each original train i, called the *scheduled version* of i.

Note that whenever the scheduled version of i is a canceled train, it means that the suggested timetable solution does not contain the train corresponding to original train i. Since a canceled train is not associated with a route, departure and arrival times will never be assigned to it, and "scheduled" in this case represents being "selected" rather than actually being scheduled.

5.4. Scheduled delay

The scheduled delay δ^i of train i is the difference between the arrival time of the scheduled train version of i, and the scheduled arrival time of the original train in the original timetable (the published train plan), at the end location.

Note that the scheduled delay can be negative. In practice this happens when the original train is associated with a lot of pathing time whereas the alternative train is not, or on rare occasions when the alternative route is shorter or takes less time to traverse than the original route.

5.5. Successful redirection

We assume that all scheduled original versions of trains in the suggested timetable solution are acceptable to the RUs, provided that the suggested timetable solution respects all safety rules and headway requirements, unless an exception has explicitly been defined.

Redirected trains are not automatically acceptable, however, and we define successful redirection of an original train i as follows:

- There is a train version i_n ($n > 0$) that is not in conflict with any track possession or train version scheduled in the suggested timetable solution
- The scheduled alternative train path of the redirected train is acceptable to the RU
- No train paths for original trains on parts of the network not directly affected by track possessions were modified to accommodate for the alternative train path of the redirected train

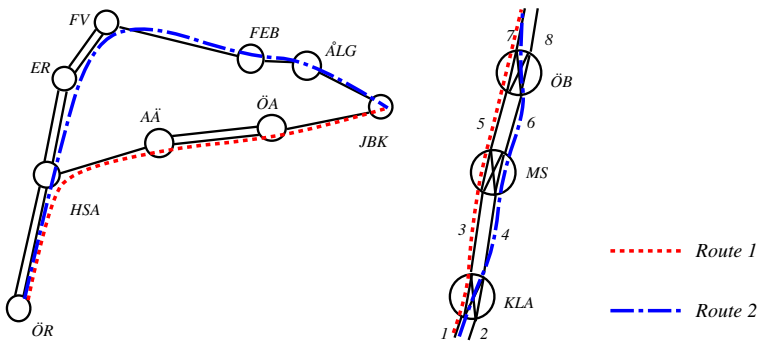

Fig. 2. Two examples of alternative routes (assuming Route 1 and Route 2 have the same direction). Note that both routes in the right picture pass the exact same locations. The reason they are alternative routes and not identical routes is because their link usages differ.

Note that the last requirement does not need to be met if the affected RUs explicitly agree to change their train paths in question.

6. The implementation

The mathematical formulation of a timetable given in Section 4 describes a very general timetable, where trains might be scheduled at any times relative to each other as long as the resource constraints of the infrastructure are respected. To be useful in practice, the implementation of the model includes measures to increase the chances that the suggested timetable solution will be meaningful for the problem at hand. The key features of our implementation will be described in this section.

6.1. Tolerance threshold

The *tolerance threshold* for an alternative train specifies the upper bound for the scheduled delay of the train versions based on it.

To keep running times within reasonable limits without adding complexity, we also forbid a train version from departing earlier from its start location than its corresponding original train in the original timetable. In this way, the tolerance threshold also effectively gives an upper bound for the maximum prolongation of running time for the train.

While the assumption that a rescheduled train must not depart earlier than in the original timetable might not always be true, for most cases it is a reasonable requirement. Passengers can wait for a train at a station, but will risk missing the train if it departs earlier than in the usual timetable.

6.2. Bounds

For the trains, the bounds for the departure and arrival times, and durations of dwell times at locations, are controlled by the following parameters:

- Stop duration requirements dictating how much the individual duration of a stop at a location can be prolonged
- Arrival requirements dictating what the latest arrival time can be for a particular train at a particular location
- Departure requirements dictating what the earliest departure time can be for a particular train from a particular location

Additional departure requirements for the start locations of the alternative trains enforce their departure times to be equal to, or bigger than, the corresponding original departure times of their respective original trains. Arrival requirements are then imposed such that they ensure running times of the alternative trains that do not exceed the scheduled running times of the original trains plus the respective train's tolerance threshold.

6.3. Allowing conflicts

Given a particular set of restrictions on the arrival and departure times of trains, it is likely that a conflict-free schedule does not exist. Allowing conflicts and highlighting them can provide the user with valuable feedback on what measures might be necessary in order to get a conflict-free solution in the next iteration. He/she might decide to soften some constraints, e.g. by loosening arrival and departure requirements, allowing some trains to be canceled, etc. A new optimization can then be performed. This iterative process stops when the solution is conflict-free, or at least contains only acceptable conflicts.

To allow a conflict in the solution between two trains i and j on link λ, we use a binary *conflict variable* C_λ^{ij} and once again a big-M formulation. M is big enough to achieve the effect that the conflict can be ignored when C_λ^{ij} is equal to 1.

Thus, the link equations (2) can be extended with a conflict variable like this:

$$\begin{aligned} d_\lambda^j - d_\lambda^i - t_\lambda^i + M\left(1 - x_\lambda^{ij}\right) + Mc_\lambda^{ij} &\geq 0 \\ d_\lambda^i - d_\lambda^j - t_\lambda^j + Mx_\lambda^{ij} + Mc_\lambda^{ij} &\geq 0 \end{aligned}$$

Conflict variables are used analogously in equation group (3).

To prevent the solution from having more conflicts than necessary, the objective function minimizes the number of conflicts.

Note that even a published train plan is usually not entirely conflict-free, but allows resource conflicts under certain circumstances. For instance, even if a headway of three minutes is usually required for the separation of trains on double-track lines, trains might occasionally be scheduled using a slightly smaller headway. Our implementation allows such original conflicts to remain in the solution.

6.4. Train versions

One of the most important outcomes of the optimization is which version of each train that will be scheduled, or if it will be canceled. Obviously, there is no need to resolve conflicts involving train versions that end up being discarded in favor of another version of the same train.

For original train i we introduce n versions: the original version i_0, and one version for each alternative train that we want to evaluate as an option to i. When the binary *version variable* v^{i_n} in equals 0, train version i_n is in the suggested timetable solution. Thus we require

$$v^{i_0} + v^{i_1} \ldots + v^{i_{n-1}} = n - 1$$

for all trains versions of the same train. We use version variables in equations in the same way as we use conflict variables.

6.5. Cancelable trains

As an option to alternative trains for an original train, or as a complement, it is possible to allow for a train to be canceled altogether. If a train is canceled, it means that none if its versions need to have their conflicts resolved. We introduce cancellation variables and use them in the same way as we use conflict and version variables.

6.6. Alternative track possessions

Every track possession is associated with either a specific desired start time, or a set of options for desired start times. The scheduled start time d^p of a track possession in the suggested timetable solution is allowed to deviate from the desired time with up to the size of a given input parameter.

If there are options for the desired start time, each one is seen as a version of the track possession. Which version that will be scheduled, and therefore needs to have its conflicts resolved, is decided with version variables in analogy with the way train versions are selected during optimization.

7. Generating good solutions

There are potentially many mathematically feasible solutions to the timetable problem defined in Sections 4 and 5. In order to be able to evaluate the suggested timetable solution and its applica-

bility to the real-world situation, the planner might want to optimize, i.e. minimize or maximize, certain timetable properties.

7.1. Relevant properties

We have identified four general properties that we find especially relevant in the context of simultaneous train and track possession scheduling, and we discuss them in this section.

The four general properties for which our model is especially well suited for optimizing are:

- Soundness
- Disruptiveness
- Cost
- Delay

A conflict-free timetable is considered a sound timetable, while the number of canceled trains for a given set of track possessions indicates how disruptive to the traffic the possession is. Cost is the monetary cost of applying the solution, i.e. giving it to the dispatching center and running the traffic accordingly. Delay is the sum of train delays in the suggested timetable solution compared with the original trains in the published timetable.

Note that ensuring soundness can be done either by forbidding resource conflicts, or making sure that the presence of resource conflicts is minimized. The advantage of the second approach is that there will always exist a solution to the problem, albeit not a sound one. Remember that a conflict in the model does not necessarily correspond to a real-world resource conflict (see Section 6.3), and there is of course no need to minimize such conflicts.

The model can minimize the total monetary cost of the solution, provided that accurate data exist for the cost of canceling, redirecting or delaying individual trains, and assuming that the costs grow linearly. On a deregulated market, however, such data does not exist since there is no way for the IM to get correct data without asking the RUs, and the RUs unfortunately have no incentive to provide the IM with accurate figures (Klabes, 2010).

7.2. Relative importance of objectives

By grouping the relevant properties and calibrating the costs for them in such a way that the more important objectives are surely fulfilled before the next most important objectives, the suggested timetable solution will reflect their relative importance.

In general, and provided that trains can be categorized as cancelable and not cancelable, the first and most important objective is to find a sound timetable, which is achieved by minimizing the presence of conflicts.

Even if canceled trains in a suggested timetable solution do not automatically disqualify a solution from being a feasible option for a new plan, in general we can assume that cancellations are highly undesirable. The second objective is therefore to minimize the number of canceled trains.

Last, i.e., only if it does not prevent performing well in the other two objectives, the objective is to minimize the sum of the scheduled delays.

8. Input data considerations

Various implementation choices have naturally been highly influenced by what kind of data has been available, and what we expect to be available in the future in terms of data. Also, the intended main application of the model has influenced everything from pure modeling choices to practical coding decisions. To provide context without going into detail, the expected input data are briefly described here.

The main input is an existing train plan. The model as such does not care whether this plan is a draft or the published, yearly plan, but for the sake of the problem discussed in this paper, we assume that an existing, yearly train plan is used as input, from which the day in question has been rolled out for a part of the network of limited size, so that precisely the trains that run on that particular day and in that geographical area are included. Arrival and departure times for all trains at all locations are specified down to the second in this plan.

When it comes to input data for track possessions, and in particular data concerning what the RUs would do if they face the option of either canceling trains or redirecting them due to track possessions, we have to rely on assumptions rather than on real data. If the model would be applied to a real case, the input would be real applications for alternative train paths, and they would automatically reflect the wishes of the RUs. But as long as we are only testing the model, we will assume the following:

- There are explicit alternative route definitions for all trains that the RUs might need to redirect in the given problem instance, from which alternative trains can be generated
- There exist clearcut definitions for what would be acceptable to the RU in terms of running times and other properties of the alternative train paths in the suggested timetable solution

For every track possession that is about to be scheduled, we assume that we have an exhaustive list of links that will be closed during the track possession, the duration of the track possession, and time windows for when the possession can be scheduled to start.

The time windows represent the final outcome of any crew availability restrictions and all other constraints that were uncovered by the planner when he/she analyzed the situation.

In order to solve the problem and find a feasible schedule, train versions for all original trains that overlap with any of the time windows of at least one of the track possessions, are generated, including versions based on any suitable alternative routes that have been identified.

9. Test scenarios

We have evaluated our model on several typical scenarios that represent relevant use cases. In this section, we describe two such scenarios, discuss their relevance, and briefly present the results of applying our model to 94 trains, constituting one day of traffic in April 2012 in the part of the Swedish network depicted in Fig. 3.

9.1. Scenario 1

Maintenance on a double-track section can often be performed on one track at a time, keeping one track available to the trains while the work is being done. If the work can be performed in a safe manner and does not take very much longer because the track section was not completely shut down, this is undoubtedly preferred to canceling all the traffic on the line for the duration of the job.

There are often a lot of trains scheduled on a double-track line, especially during peak hours. The capacity on a line decreases significantly when it must suddenly be treated as a single-track line instead of as having the usual double-track property. In this context, our model could e.g. be used to provide the IM with an automatic analysis of how different parameters would affect the traffic.

The affected infrastructure was one of the tracks on the double-track section between Kumla and Örebro Södra. The duration of the track possession was fixed to be 14 h, but the model was free to

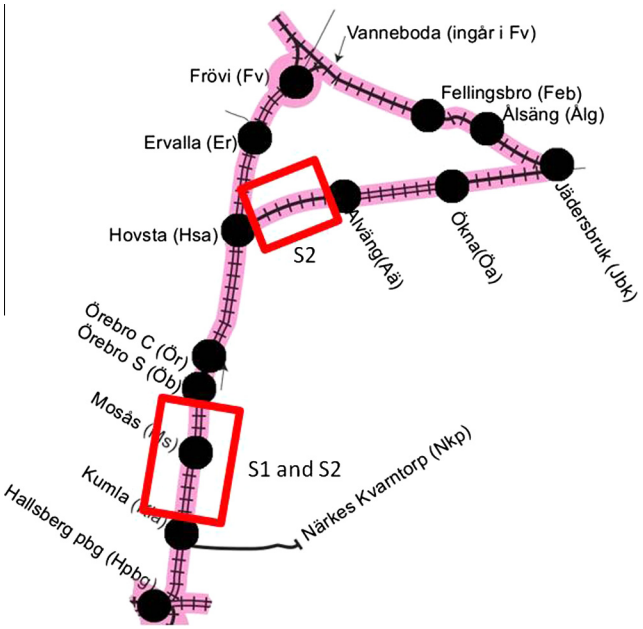

Fig. 3. The part of the network used for the two test scenarios. The sections affected by the track possessions in Scenario 1 (S1) and Scenario 2 (S2) are marked with rectangles.

find the most suitable start time within a time window of 2 h centered around an arbitrarily chosen start time at 04:00 a.m.

We defined alternative routes for all original trains that used the links that would be closed during the track possession at a time when the track possession might affect them, i.e., between 3 a.m. and 7 p.m. The alternative routes all passed the same locations as the original routes, but used the right track of the double-track section between Kumla and Örebro instead of the usual left track (in Sweden, the default track for traveling on a double-track section is the left track). For each such alternative route and original train, we created an alternative train, and for each original and alternative train, we created a train version. The number of alternative trains was 30.

The stopping behavior of the alternative trains were similar to that of their corresponding original trains', with the exception that they were all given the option to stop and wait for other trains at the locations Kumla and Örebro even if the original trains did not stop there.

We tested this scenario for two different settings. In Setting 1, only the trains in the direction directly affected by the closed track were allowed to be redirected or canceled. No scheduled delays were allowed for the original versions, whereas alternative trains were allowed a scheduled delay up to the size of a specified common tolerance threshold. Setting 2 allowed a scheduled delay for all trains, regardless of direction, within the bounds of the given tolerance threshold. In all other respects, the two settings were the same, and all trains directly affected by a track possession were considered to be cancelable.

The first objective in both settings was to minimize the number of canceled trains, and the second objective was to minimize the sum of the scheduled delays. All trains, regardless of whether they were redirected or not, were required to depart either at the same time as the corresponding trains in the published train plan, or later.

In the diagrams of Figs. 4 and 5, the correlations between the tolerance threshold and the number of canceled trains and the total scheduled delay are shown. With a tolerance threshold of 840 s (14 min), all trains could be scheduled in Setting 2, whereas eight trains would still have to be canceled in Setting 1. Only with a tolerance threshold of 2820 s (47 min), could all trains in Setting 1 finally be scheduled. This shows the importance of planning ahead to be able to use the remaining capacity efficiently in the event of a track possession on a double-track line.

The optimization was performed with CPLEX 12.2 on a Lenovo ThinkPad T430s, with Intel processor Intel(R) Core(TM) i7 2.90 GHz, under Ubuntu Linux. The CPU time did not exceed 17 s for any combination of parameters, and for Setting 1 it never exceeded 1 s. Generating the problem itself (building the problem and writing the equations to a file on a suitable format for CPLEX) never took more than 1 min. These CPU times would be acceptable to a user of a tool implementing our model.

9.2. Scenario 2

A situation that occurs frequently is when maintenance has to be done on a single-track line section after the yearly train plan has been published, and it cannot be scheduled between two trains. Applying our model in such a situation can help find a new schedule that minimizes the disturbance to the traffic.

We assume that there are alternative trains for all original trains for which suitable alternative routes exist, and that all alternative trains are associated with a tolerance threshold. Furthermore, the start times of each track possession that needs to be scheduled has to lie in precisely one of possibly many explicitly specified time windows.

We defined the problem as finding the best schedule for two fictitious track possessions in the input timetable, given three time windows for each of the track possessions. In our test scenario, the same tolerance values were used for all alternative trains. All trains that were directly affected by a track possession were considered to be cancelable.

One of the track possessions closed the single-track line between Hovsta and Alväng (see Fig. 3), a track section with fairly low volume of traffic, and had a fixed duration of 7 h. The other track possession closed one of the tracks on the double-track section between Kumla and Örebro Södra. We performed separate tests using 10 different values for the duration of the second track possession (0, 1, 2, ..., 9 h) and observed how these affected selected parameters for tolerance thresholds of 0, 5, 10 and 15 min. The number of train versions generated in addition to the original 94 versions ranged between 41 and 61 for this scenario.

Fig. 6 shows a diagram where the track possession duration is plotted against the number of canceled trains for four different tolerance threshold values. Such a diagram can be used to quickly see how big the tolerance threshold needs to be in order to make room for a track possession of a certain length. E.g., Fig. 6 shows that a tolerance threshold of at least 15 min (900 s) would enable all trains to be scheduled as long as the duration of the second track possession would be 3 h or less. Likewise, Fig. 6 reveals that a

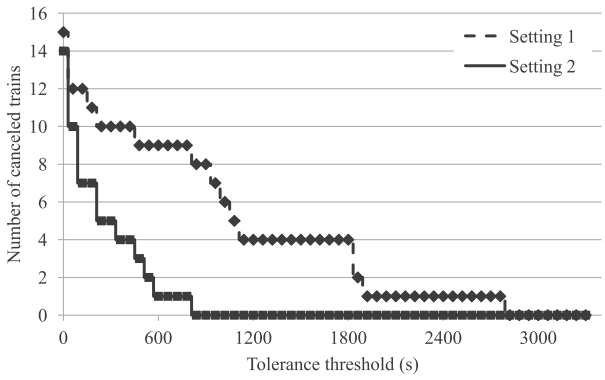

Fig. 4. The correlation between the tolerance threshold and the number of canceled trains for the two settings in Scenario 1.

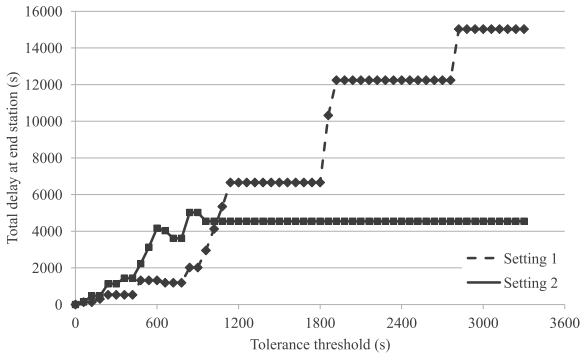

Fig. 5. The total scheduled delay plotted against the tolerance threshold for the two settings in Scenario 1.

tolerance threshold of 15 min (900 s) is not enough to schedule the second track possession without needing to cancel at least two trains if the duration of the second track possession is 5 h or more, and that a tolerance threshold of 10 min (600 s) would give the same number of canceled trains in that particular case.

Table 1 displays how many of the original 94 trains that had to be canceled for all 10 different track possession durations when the tolerance threshold was 5 min. The number of trains which were realized by train versions based on alternative routes is presented as Changed routes. The sum of the delays of all trains is given as Total delay. The CPU time column gives the time CPLEX needed to solve the optimization problem on the same computer as the one used for Scenario 1.

9.3. Observations

Choosing where to draw the line between what is in the model and what is not, is in general a tough task. The guiding principle is that there has to be more to be gained by considering e.g. a bigger geographical area than we lose in complexity and energy spent on modeling and solving the problem, but this is not always easy to predict in advance.

In our case, canceling trains from one particular part of the network could mean that they can still be considered for redirection to parts of the network that are not explicitly in the model. For Scenario 1, it would have been possible to consider a larger part of the infrastructure to be able to introduce more alternative routes as options in the same way as in Scenario 2, and evaluate them all simultaneously. However, depending on how the result of the optimization is going to be used, it might be sufficient to consider

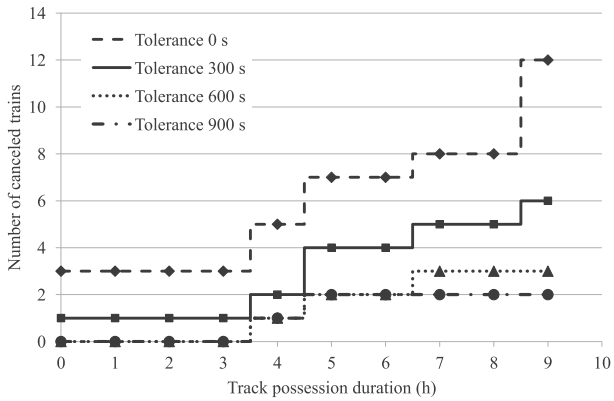

Fig. 6. The track possession duration plotted against the number of canceled trains in Scenario 2, for four different tolerance threshold values.

Table 1
Results for different track possession durations (tolerance threshold 5 min).

Duration (h)	Total delay (s)	Canceled trains (#)	Changed routes (#)	CPU time (s)
0	343	1	4	0.05
1	343	1	7	0.07
2	343	1	6	0.09
3	343	1	7	0.08
4	570	2	9	0.09
5	481	4	11	0.11
6	481	4	12	0.11
7	514	5	12	0.15
8	514	5	11	0.12
9	541	6	13	0.10

only the more limited part of the network that we used for Scenario 1.

We believe that both scenarios represent typical situations where an optimization model of the kind we have developed is particularly useful in practice. Already with limited knowledge about what the RUs would be willing to accept, the model is able to give a clear indication of how severe a situation is. Such an initial analysis would be very easy and fast to perform, and it could be used to decide when a complete re-planning process is actually called for. This impression was confirmed when these ideas were presented to a group of relevant actors in 2012 (Trafikverket, 2012).

10. Summary and future research

We have developed a MIP model that, given an existing timetable and a fixed set of track possessions, reschedules trains in a way that disturbs the flow of traffic as little as possible. In addition to putting the model into context and describing it from a mathematical point of view, we have presented the results of applying it on a couple of typical scenarios based on real data.

In principle, the planner has three different tools with which to make room for track possessions: cancellations, redirections, and delays. We have not yet investigated the relation between these three, or how they should be weighed against each other. Making room for possessions in a way that minimizes the overall monetary cost of the traffic disturbance and the track possession itself is the most obvious goal to strive for, but it is not meaningful on a deregulated market since correct cost data are impossible for the IM to obtain. In the light of this, how to define a good solution is an important topic for future research, and involves socio-economic considerations as well as mathematical modeling challenges.

The test scenarios in Section 9 show that the expressive power of the model is sufficient for small real-world cases. Minimizing three different objectives in one single step might however quickly lead to long computation times with increasing problem sizes. One of the challenges that remains before the model can be useful in practice is to make sure that the model can handle larger problem instances. The next step for us therefore involves finding an objective function that both scales well and captures the problem characteristics better than our current approach.

Acknowledgments

The project has been funded by Trafikverket (the Swedish Transport Administration), Grant TRV 2010/25229. We would like to take the opportunity to thank Hans Dahlberg, Per Edholm and Jakob Fritzell at Trafikverket for contributing with their expert knowledge, and the anonymous reviewers for their valuable comments that helped to improve this paper.

Appendix A

The railway network is modeled by geographical locations (also called link nodes) $g \in G$ and links $\lambda \in \Lambda$. Each link λ represents a unique track section connecting two geographical locations. The set of links Λ comprises two disjoint sets Λ_S and Λ_M. A link connecting two link nodes g_1 and g_2 is a single-track link $\lambda \in \Lambda_S$ iff no other link in Λ connects the same two link nodes, and a multi-track link $\lambda \in \Lambda_M$ otherwise. $G_L \in G$ is the set of geographical locations that do not allow for the train order to change (e.g. block signals).

The traffic is modeled by an ordered set of trains I comprising two disjoint sets of original trains I_0 and train versions of original trains I_V. The set of train versions derived from original train $i_0 \in I_0$ is denoted $I(i_0)$.

All trains $i \in I$ are associated with a unique route represented by an ordered subset of links, $\Lambda(\mathrm{i}) \subseteq \Lambda$. For any link λ in a route $\Lambda(\mathrm{i})$, let $\lambda+1$ denote the link following λ in the route, if it exists. The geographical location connecting two links λ and $\lambda+1$ is given by $g(\lambda, \lambda+1)$. If routes (i.e. ordered links) for two different trains are discussed simultaneously, superscripts are used to denote which link order that belongs to which train route. The first link of the route of train i is denoted λ_0^i, and the last link λ_f^i. Likewise, let $G(i)$ denote the set of geographical locations that train i visits.

Depending on which direction train i is traveling in, it will reach one of the locations of a link first, $g_1(i,\lambda)$, and the other one later, $g_2(i,\lambda)$. The time it takes for train i to traverse a link λ is called the link trip time, t_λ^i. The time when train i arrives to a geographical location g is called the arrival time, a_g^i, and the time when train i enters a link λ is called the departure time, d_λ^i.

Trains may stop at geographical locations and the dwell time for train i at location g is denoted w_g^i. Based on physical and commercial properties of a geographical location, and given as input, there is a set of geographical locations $G_S(i) \subseteq G$ where train i must stop, a set of locations $G_V(i) \subseteq G$ where train i may stop, and a set of locations $G_N(i) \subseteq G$ where train i must not stop. For geographical locations $g \in G_N(i)$, the dwell time for train i is not a variable but fixed at zero, $w_g^i = 0$. For geographical locations where the train has to stop there is a minimum dwell time denoted $\underline{w}_g^i$. For $g \in G_V(i)$, a binary variable z_g^i is used to denote if train i stops at location g or not.

The original timetable is defined by the set of original trains I_0 and their respective original arrival and departure times. Whether a train $i \in I$ is included in the optimized timetable or not is encoded by the binary variable v_i, which equals 0 for all trains to be included. Further, a pre-determined set of original trains may be canceled, $i_0 \in I_C$. Cancellations of trains are modeled using binary variables c^{i_0} which equals 0 if train i_0 is canceled. Iff an original train $i_0 \in I_0$ is not canceled, either the original train itself or precisely one if its train versions $i \in I(i_0)$ is included in the optimized plan.

Geographical locations can only host a certain number of trains simultaneously, i.e. location g has capacity $n(g)$. In Aronsson et al. (2009), a method where the problem of scheduling trains on stations is modeled as a graph is described, with nodes representing trains and arcs potential concurrent use of a station. Further, Aronsson et al. (2009) shows how to find a set of minimal sub-cliques $M \in S(g)$ for each geographical location g. Each such minimal sub-clique M is a set of trains that must not all be at the station simultaneously, but that would be schedulable on the station tracks if at least one of the trains in M is absent.

For two trains i and j, $i<j$, that may use a geographical location g at the same time, a binary variable u_g^{ij} is introduced to encode whether the two trains are simultaneously using g or not. The station capacity $n(g)$, given as input, gives an upper limit of how many trains that can reside at location g simultaneously. As it is assumed that all trains fit on all station tracks, and require one and only one track, all sub-cliques for a geographical location g will be of size $n(g)+1$ and have $\binom{n(g)+1}{2}$ arcs. By ensuring that for every minimal sub-clique at least one of the potential overlaps represented by the arcs in the clique is not realized, the station capacity constraints are respected. That is, the sum of binary variables for simultaneous use of a station has to be less than or equal to $\binom{n(g)+1}{2} - 1$ for trains in a minimal sub-clique. A binary conflict variable c_g^{ij} takes value 1 if trains i and j are present at location g simultaneously even if this breaks the capacity constraint of g. A pre-determined, possibly empty, set of allowed conflicts (i, j) at geographical location g are given by $C(g)$.

Further, let U_g denote the least time interval required between the arrivals of two trains at location g. A conflict variable q_g^{ij} encodes if this least interval is not maintained, and the sequencing binary variable x_g^{ij} takes the value of 1 if i arrives at location g before train j. Once again, only a pre-determined set of conflicts against the arrival separation rule is allowed at a geographical location g, and the conflicts are given by $(i,j) \in C_U(g)$. Further, for a set of stations $g \in G_p$ this constraint is ignored if both trains are scheduled to stop at g. This is modeled using a variable p_g^{ij}, which may take a value of 1 if both trains i and j stop at g, and big-M constraints.

There are bounds on all arrival and departure times, link trip times and dwell-times. The notation v and $\overline{v}$ is used to denote the lower and upper bound of a variable v.

The minimum link trip time is governed by the stop pattern of the train, i.e. whether the train stops at the link nodes connected by the link or not. The four link stop patterns are *FF* (full-speed at both link nodes), *SF* (stop at first link node, full speed at second), *FS* (full speed at first link node, stop at second) and *SS* (stop at both link nodes). The minimum link trip time for a certain stop pattern is denoted $T_{\lambda,X}^i$ where $X \in \{FF, FS, SF, SS\}$. To easily obtain by how much the trip times have been prolonged during the optimization (i.e. how much of the trip time is so called pathing time), slack variables s_λ^i are introduced.

We assume that there is a desired time α^i for the arrival at the end station for every train $i \in I$ and let δ^i denote the difference between the arrival time in the optimized timetable and this desired arrival time.

Trains that need the same link must be sequenced on the link and separated adequately in time. The binary variable x_λ^{ij} is 1 if train i enters link λ before train j. The set $I(\lambda)$ is an ordered set of trains that traverse link λ. Further, $O(i,\lambda)$ is the set of trains j that, according to the arrival and departure time bounds, could overlap with train i on a link λ, and where $j>i$ according to the order of I.

For trains traveling in opposite directions the first train must have finished the link trip before the second train can enter the link. This is also true for trains traveling in the same direction on a single-track link, $\lambda \in \Lambda_s$. However, for trains traveling in the same direction on a multi-track link $\lambda \in \Lambda_M$, we assume that it suffices to separate the two trains with a headway. This headway depends on the link and which train i that enters the link first, and it is denoted HW_λ^i. A binary conflict variable c_λ^{ij} takes value 1 if trains i and j are not separated adequately on link λ. Just like with conflicts at geographical locations, a pre-determined, possibly empty, set of allowed conflicts (i, j) at link λ are given by $C(\lambda)$.

The model includes the possibility to require a time difference r_g^{ij} between the arrival of train i and the departure of train j at link node g. These constraints can be used to e.g. ensure that there is enough time for passengers, or a train driver, to move from one train to another. The set of trains that are required to depart later or at the same time as the arrival time of train i at a geographical location g is given by $A_1(i,g)$, and those that have to depart earlier or at the same time is given by $A_2(i,g)$.

Finally, we let P denote the set of track possessions. Just like for trains, we identify a set of "original track possessions" $P_0 \subseteq P$ and let each original track possession $p_0 \in P_0$ have a set of versions, $P(p_0) \subset P$. The start time d^p of a track possession p is associated with an interval during which it must start, $\underline{d}^p \le d^p \le \overline{d}^p$, and a duration t^p. Also, the links affected by a track possession p are given by $\Lambda(p)$. A binary sequencing variable x_λ^{ip} is used to capture if train i travels on a link λ before or after track possession p. $O(p,\lambda)$ is used to denote the set of trains that may be affected by a track possession p on link λ, meaning that these trains overlap with the possession according to the respective bounds of start and end times of the train traversals and possession in question.

The objective function is threefold. Variables are included to minimize the number of resource conflicts and canceled trains, and to make sure that train paths close to the desired ones are generated. The cost of canceling train i is denoted C_{c_i}. The cost of a link conflict between trains i and j is denoted $C_{c_\lambda^{ij}}$, a location conflict $C_{c_g^{ij}}$, and a violation against the time needed between the arrivals of two trains at a geographical point $C_{U_g^{ij}}$. Finally, train version $i \in I_v$ carries a cost C_{v^i} if included in the optimized timetable, and the difference δ^i between the arrival time of a train at its final location in the optimized timetable and the desired arrival time at that location is included.

The constrains have been divided into categories, and headings are included to facilitate reading. Note that, as stated above, there is a set of allowed conflicts for every conflict type, and conflict variables are only defined for allowed conflicts. This means that whenever there is a constraint group where some conflicts may be allowed there should be two sets of similar constraints, one including conflict variables to allow for conflicts, and one without conflict variables. Only the first set of constraints is included below, and for all situations where a conflict is not allowed the conflict variable should simply be omitted. The same is true for constraints with p_g^{ij}, i.e. the variable is included for relevant combinations of trains and locations, and omitted otherwise. Also, big-M constraints are used multiple times in the model, where M is a number large enough to dominate the constraint.

$$\min \sum_{\lambda\in\Lambda}\sum_{(i,j)\in C(\lambda)} C_{c_\lambda^{ij}} c_\lambda^{ij} + \sum_{i\in I}\delta^i + \sum_{i\in I} C_{v^i}(1-v^i) + \sum_{g\in G}\sum_{(i,j)\in C(g)} C_{c_g^{ij}} c_g^{ij} + \sum_{g\in G}\sum_{(i,j)\in C_U(g)} C_{U_g^{ij}} q_g^{ij} + \sum_{i\in I_C} C_{c^i}(1-c^i)$$

s.t

Bounds on departure times, arrival times and total travel time.

$$\begin{aligned}
&\underline{d}_\lambda^i \le d_\lambda^i \le \overline{d}_\lambda^i && i\in I,\ \lambda\in\Lambda(i)\\
&\underline{a}_{g_2(i,\lambda)}^i \le a_{g_2(i,\lambda)}^i \le \overline{a}_{g_2(i,\lambda)}^i && i\in I, \lambda\in\Lambda(i)\\
&a_{g_2(i,\lambda_f^i)}^i - \delta^i - Mv^i \le \alpha^i && i\in I\\
&a_{g_2(i,\lambda_f^i)}^i + 10\delta^i + Mv^i \ge \alpha^i && i\in I\\
&\delta_i \ge 0 && i\in I
\end{aligned}$$

Trip consistency constraints.

$$\begin{aligned}
&d_\lambda^i + t_\lambda^i = a_{g_2(i,\lambda)}^i && i\in I,\ \lambda\in\Lambda(i)\\
&a_{g_1(i,\lambda)}^i + w_{g_1(i,\lambda)}^i = d_\lambda^i && i\in I,\ \lambda\in\Lambda(i)
\end{aligned}$$

Stop patterns and their effects on the trip times.

$$\begin{aligned}
&w_g^i = 0 && i\in I,\ g\in G_N(i)\\
&\underline{w}_g^i \le w_g^i \le \overline{w}_g^i && i\in I,\ g\in G_S(i)\\
&0 \le w_g^i \le \overline{w}_g^i && i\in I,\ g\in G_V(i)\\
&w_g^i - Mz_g^i \le 0 && i\in I,\ g\in G_V(i)\\
&t_\lambda^i - s_\lambda^i = T_{\lambda,FF}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_N(i)\wedge g_2(i,l)\in G_N(i)\}\\
&t_\lambda^i - s_\lambda^i = T_{\lambda,FS}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_N(i)\wedge g_2(i,l)\in G_S(i)\}\\
&t_\lambda^i - s_\lambda^i = T_{\lambda,SF}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_S(i)\wedge g_2(i,l)\in G_N(i)\}\\
&t_\lambda^i - s_\lambda^i = T_{\lambda,SS}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_S(i)\wedge g_2(i,l)\in G_S(i)\}\\
&t_\lambda^i - s_\lambda^i + \left(T_{\lambda,SF}^i - T_{\lambda,SS}^i\right) z_{g_2(i,\lambda)}^i = T_{\lambda,SF}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_S(i)\wedge g_2(i,l)\in G_V(i)\}\\
&t_\lambda^i - s_\lambda^i + \left(T_{\lambda,FF}^i - T_{\lambda,FS}^i\right) z_{g_2(i,\lambda)}^i = T_{\lambda,FF}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_N(i)\wedge g_2(i,l)\in G_V(i)\}\\
&t_\lambda^i - s_\lambda^i + \left(T_{\lambda,FS}^i - T_{\lambda,SS}^i\right) z_{g_1(i,\lambda)}^i = T_{\lambda,FS}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_S(i)\}\\
&t_\lambda^i - s_\lambda^i + \left(T_{\lambda,FF}^i - T_{\lambda,SF}^i\right) z_{g_1(i,\lambda)}^i = T_{\lambda,FF}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_N(i)\}\\
&f_{\lambda,FF}^i + f_{\lambda,SF}^i + f_{\lambda,FS}^i + f_{\lambda,SS}^i = 1 && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_V(i)\}\\
&f_{\lambda,FF}^i + z_{g_1(i,\lambda)}^i + z_{g_2(i,\lambda)}^i \ge 1 && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_V(i)\}\\
&f_{\lambda,SF}^i - z_{g_1(i,\lambda)}^i + z_{g_2(i,\lambda)}^i \ge 0 && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_V(i)\}\\
&f_{\lambda,FS}^i + z_{g_1(i,\lambda)}^i - z_{g_2(i,\lambda)}^i \ge 0 && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_V(i)\}\\
&z_{g_1(i,\lambda)}^i + z_{g_2(i,\lambda)}^i - f_{\lambda,SS}^i \le 1 && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_V(i)\}\\
&t_\lambda^i - s_\lambda^i = T_{\lambda,FF}^i f_{\lambda,FF}^i + T_{\lambda,SF}^i f_{\lambda,SF}^i + T_{\lambda,FS}^i f_{\lambda,FS}^i + T_{\lambda,SS}^i f_{\lambda,SS}^i && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_V(i)\}\\
&f_{\lambda,X}^i \ge 0 && i\in I,\ \lambda\in\{l\in\Lambda(i)\mid g_1(i,l)\in G_V(i)\wedge g_2(i,l)\in G_V(i)\},\ X\in\{FF,FS,SF,SS\}\\
&\overline{t}_\lambda^i \ge t_\lambda^i && i\in I,\ \lambda\in\Lambda(i)\\
&0 \le s_\lambda^i && i\in I,\ \lambda\in\Lambda(i)
\end{aligned}$$

Selecting which train versions to include in the optimized timetable.

$$v^{i_0} + c^{i_0} + \sum_{j \in I(i_0)} v^j = |\, I(i_0) \,| + 1 \quad i_0 \in I_0$$

Sequencing trains on single-track links and trains traveling in opposite direction on multi-track links.

$$d^j_\lambda - d^i_\lambda - t^i_\lambda + M\left(1 - x^{ij}_\lambda\right) + M\left(c^{ij}_\lambda + v^i + v^j\right) \geq 0 \quad \lambda \in \Lambda_S,\ i \in I(\lambda),\ j \in O(i,\lambda) \text{ and } \lambda \in \Lambda_M, i \in I(\lambda),\ j \in \{k \in O(i,\lambda) \mid g_1(i,\lambda) \neq g_1(k,\lambda)\}$$

$$d^i_\lambda - d^j_\lambda - t^j_\lambda + Mx^{ij}_\lambda + M\left(c^{ij}_\lambda + v^i + v^j\right) \geq 0 \quad \lambda \in \Lambda_S,\ i \in I(\lambda),\ j \in O(i,\lambda) \text{ and } \lambda \in \Lambda_M, i \in I(\lambda),\ j \in \{k \in O(i,\lambda) \mid g_1(i,\lambda) \neq g_1(k,\lambda)\}$$

Sequencing trains traveling in the same direction on multi-track links.

$$d^j_\lambda - d^i_\lambda - HW^i_\lambda + M(1 - x^{ij}_\lambda) + M\left(c^{ij}_\lambda + v^i + v^j\right) \geq 0 \quad \lambda \in \Lambda_M,\ i \in I(\lambda),\ j \in \{k \in O(i,\lambda) \mid g_1(i,\lambda) = g_1(k,\lambda)\}$$

$$d^i_\lambda - d^j_\lambda - HW^j_\lambda + Mx^{ij}_\lambda + M\left(c^{ij}_\lambda + v^i + v^j\right) \geq 0 \quad \lambda \in \Lambda_M,\ i \in I(\lambda),\ j \in \{k \in O(i,\lambda) \mid g_1(i,\lambda) = g_1(k,\lambda)\}$$

$$a^j_{g_2(i,\lambda)} - a^i_{g_2(i,\lambda)} - HW^i_\lambda + M\left(1 - x^{ij}_\lambda\right) + M\left(c^{ij}_\lambda + v^i + v^j\right) \geq 0 \quad \lambda \in \Lambda_M,\ i \in I(\lambda),\ j \in \{k \in O(i,\lambda) \mid g_1(i,\lambda) = g_1(k,\lambda)\}$$

$$a^i_{g_2(i,\lambda)} - a^j_{g_2(i,\lambda)} - HW^j_\lambda + Mx^{ij}_\lambda + M\left(c^{ij}_\lambda + v^i + v^j\right) \geq 0 \quad \lambda \in \Lambda_M,\ i \in I(\lambda),\ j \in \{k \in O(i,\lambda) \mid g_1(i,\lambda) = g_1(k,\lambda)\}$$

Train order at geographical locations.

$$2c^{ij}_{\lambda_k} + x^{ij}_{\lambda_k} + 2c^{ij}_{\lambda_l} - x^{ij}_{\lambda_l} + M(v^i + v^j) \geq 0 \quad (i,j) \in \{(m,n) \mid m \in I \wedge n \in I \wedge m < n\},\ (\lambda_k, \lambda_l) \in \{(\lambda_o, \lambda_p) \mid \lambda_o \in \Lambda(i) \cap \Lambda(j) \wedge\ \lambda_p \in \Lambda(i) \cap \Lambda(j) \wedge |\, \lambda^i_o - \lambda^i_p \,| = 1 \wedge\ |\, \lambda^j_o - \lambda^j_p \,| = 1 \wedge g(\lambda_o, \lambda_p) \in G_L\}$$

$$a^j_g - a^i_g - U_g + M\left(1 - x^{ij}_g\right) + M\left(q^{ij}_g + v^i + v^j + p^{ij}_g\right) \geq 0 \quad g \in G, i \in I(g), j \in 0(i,g)$$

$$a^i_g - a^j_g - U_g + Mx^{ij}_g + M\left(q^{ij}_g + v^i + v^j + p^{ij}_g\right) \geq 0 \quad g \in G, i \in I(g), j \in O(i,g)$$

$$z^i_g - p^{ij}_g \geq 0 \quad i \in I, g \in G_P \cap G_V(i), j \in \{k \in O(i,g) \mid g \in G_V(k) \cup G_S(k)\}$$

$$z^j_g - p^{ij}_g \geq 0 \quad i \in I, g \in G_P \cap (G_V(i) \cup G_S(i)), j \in \{k \in O(i,g) \mid g \in G_V(k)\}$$

$$p^{ij}_g \geq 0 \quad i \in I, g \in G_P \cap (G_V(i) \cup G_S(i)), j \in \{k \in O(i,g) \mid g \in G_V(k) \cup G_S(k)\}$$

Capacity constraints for geographical locations. ε is a small number that ensures that trains not stopping at the station also require some capacity.

$$a^j_{g_2(i,\lambda)} - a^i_{g_2(i,\lambda)} - w^i_{g_2(i,\lambda)} + M\left(1 - x^{ij}_{g_2(i,\lambda)}\right) + M\left(u^{ij}_{g_2(i,\lambda)} + c^{ij}_{g_2(i,\lambda)} + v^i + v^j\right) \geq \varepsilon \quad i \in I,\ \lambda \in \Lambda(i),\ j \in O(i, g_2(i,\lambda))$$

$$a^i_{g_2(i,\lambda)} - a^j_{g_2(i,\lambda)} - w^j_{g_2(i,\lambda)} + Mx^{ij}_{g_2(i,\lambda)} + M\left(u^{ij}_{g_2(i,\lambda)} + c^{ij}_{g_2(i,\lambda)} + v^i + v^j\right) \geq \varepsilon \quad i \in I,\ \lambda \in \Lambda(i),\ j \in O(i, g_2(i,\lambda))$$

$$\sum_{i,j \in M: i<j} u^{ij}_g \leq \binom{n(g)+1}{2} - 1 \quad g \in G, M \in S(g)$$

Track possession constraints.

$$\underline{d}^p \leq d^p \leq \overline{d}^p \quad p \in P$$

$$v^{p_0} + \sum_{p \in P(p_0)} v^p = |\, P(p_0) \,| \quad p_0 \in P_0$$

Associations.

$$a^i_g + r^{ij}_g \leq d^j_\lambda \quad i \in I,\ g \in G(i),\ j \in A_1(i,g),\ \lambda \in \{l \in \Lambda \mid g_1(j,l) = g\}$$

$$a^i_g - r^{ij}_g \geq d^j_\lambda \quad i \in I,\ g \in G(i),\ j \in A_2(i,g),\ \lambda \in \{l \in \Lambda \mid g_1(j,l) = g\}$$

Sequencing track possessions and trains.

$$d^p - d^i_\lambda - t^i_\lambda + M(1 - x^{ip}_\lambda) + M(v^i + v^p) \geq 0 \quad p \in P,\ \lambda \in \Lambda(p),\ i \in O(p,\lambda)$$

$$d^i_\lambda - d^p - t^p + Mx^{ip}_\lambda + M(v^i + v^p) \geq 0 \quad p \in P,\ \lambda \in \Lambda(p),\ i \in O(p,\lambda)$$

Binary declarations.

$$v^i \in \{0,1\} \quad i \in I$$
$$c_g^{ij} \in \{0,1\} \quad g \in G,\ (i,j) \in C(g)$$
$$c_\lambda^{ij} \in \{0,1\} \quad \lambda \in \Lambda,\ (i,j) \in C(\lambda)$$
$$q_g^{ij} \in \{0,1\} \quad g \in G,\ (i,j) \in C_U(g)$$
$$c^i \in \{0,1\} \quad i \in I_C$$
$$x_\lambda^{ij} \in \{0,1\} \quad i \in I,\ \lambda \in \Lambda(i),\ j \in O(i,\lambda)$$
$$x_g^{ij} \in \{0,1\} \quad i \in I,\ g \in G(i),\ j \in O(i,g)$$
$$u_g^{ij} \in \{0,1\} \quad i \in I,\ g \in G(i),\ j \in O(i,g)$$
$$z_g^i \in \{0,1\} \quad i \in I,\ g \in G_V(i)$$

References

Albrecht, A.R., Panton, D.M., Lee, D.H., 2013. Rescheduling rail networks with maintenance disruptions using problem space search. Comput. Oper. Res. 40 (3), 703–712.

Andrade, A.R., Teixeira, P.F., 2012. A Bayesian model to assess rail track geometry degradation through its life-cycle. Res. Transport. Econ. 36 (1), 1–8 (Selected papers from the 12th WCTR Topic Area Transport Economics and Finance).

Aronsson, M., Bohlin, M., Kreuger, P., 2009. MILP formulations of cumulative constraints for railway scheduling – a comparative study. In: Clausen, J., Di Stefano, G. (Eds.), ATMOS 2009 – 9th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems, Dagstuhl, Germany. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, Germany.

Budai, G., Huisman, D., Dekker, R., 2004. Scheduling preventive railway maintenance activities. IEEE Int. Conf. Syst. Man Cybernet. 5, 4171–4176.

Budai-Balke, G., 2009. Operations Research Models for Scheduling Railway Infrastructure Maintenance (PhD thesis). Erasmus University, Rotterdam, The Netherlands.

Budai-Balke, G., Dekker, R., Kaymak, U., December 2009. Genetic and memetic algorithms for scheduling railway maintenance activities. Econometric Institute Report EI 2009-30, Erasmus University Rotterdam, Econometric Institute.

Cheung, B.S.N., Chow, K.P., Hui, L.C.K., Yong, A.M.K., 1999. Railway track possession assignment using constraint satisfaction. Eng. Appl. Artif. Intell. 12 (5), 599–611.

Fischetti, M., Salvagnin, D., Zanette, A., 2007. Fast approaches to robust railway timetabling. In: Liebchen, C., Ahuja, R.K., Mesa, J.A. (Eds.), ATMOS, vol. 07001 of Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, pp. 142–157.

Higgins, A., Kozan, E., 1997. Heuristic techniques for single line train scheduling. J. Heuristics 3, 43–62.

Jardine, A.K.S., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Signal Process. 20 (7), 1483–1510.

Klabes, S. Algorithmic railway capacity allocation in a competitive European railway market (PhD thesis). RWTH Aachen, Germany, 2010.

Vatn, J., 2008. Maintenance in the rail industry. In: Kobbacy, K.A.H., Prabhakar, M.D.N. (Eds.), Complex System Maintenance Handbook, Springer Series in Reliability Engineering. Springer, London, pp. 509–531.

Lake, M., Ferreira, L., 2002. Minimising the conflict between rail operations and infrastructure maintenance. In: Taylor, M. (Ed.), Transportation and Traffic Theory in the 21st Century: Proceedings of the 15th International Symposium on Transportation and Traffic Theory, Pp. Elsevier, Oxford, UK, pp. 63–80.

Larsson, D., 2004. A Study of the Track Degradation Process Related to Changes in Railway Traffic (Licentiatavhandling). Luleå tekniska universitet.

Pacciarelli, D., Pranzo, M., 2001. A tabu search algorithm for the railway scheduling problem. In: Proceedings of the 4th Metaheuristics International Conference, Porto, Portugal, pp. 159–163.

Peng, F., Ouyang, Y., 2012. Track maintenance production team scheduling in railroad networks. Transport. Res. Part B Meth 46 (10), 1474–1488.

Peng, F., Kang, S., Li, X., Ouyang, Y., Somani, K., Acharya, D., 2011. A heuristic approach to the railroad track maintenance scheduling problem. Comput. Aided Civil Infrastruct. Eng. 26 (2), 129–145.

Plu, J., Bondeux, S., Boulanger, D., Heyder, R., 2009. Application of fracture mechanics methods to rail design and maintenance. Eng. Fract. Mech. 76 (17), 2602–2611 (Special Issue on the Damage Tolerance of Railway Rails).

Pudney, P.J., Wardrop, A., 2004. Generating train plans with problem space search. In: CAPST: Ninth International Conference on Computer-Aided Scheduling of Public Transport, San Diego, USA.

Putallaz, Y., Rivier, R., 2003. Strategic maintenance and renewal policy of a railway corridor, taking into account the value of capacity. In: World Congress of Rail Research.

Simson, S., Ferreira, L., Murray, M., 2000. Rail track maintenance planning: an assessment model. Transportation Research Record: Journal of the Transportation Research Board 1713, 29–35.

The European Parliament and the Council, 2001. Directive 2001/14/EC.

Trafikverket, May 2012. Presentation of our research at a meeting for co-ordination of track possessions and trains hosted by Trafikverket.