A three-level framework for performance-based railway timetabling

Rob M.P. Goverde [a], Nikola Besinovic [a], Anne Binder [b], Valentina Cacchiani [c], Egidio Quaglietta [a], Roberto Roberti [c,d], Paolo Toth [c]

[a] Department of Transport and Planning, Delft University of Technology
[1] P.O. Box 5048, 2600 GA Delft, The Netherlands
E-mail: r.m.p.goverde@tudelft.nl
[b] Institut für Verkehrstelematik, Technische Universität Dresden, Germany
[c] DEI, University of Bologna, Italy
[d] Department of Transport, Technical University of Denmark

Abstract
The performance of railway operations depends highly on the quality of the railway timetable. In particular for dense railway networks it can be a challenge to obtain a stable robust conflict-free and energy-efficient timetable with acceptable infrastructure occupation and short travel times. This paper presents a performance-based railway timetabling framework using an integrated approach on three levels: microscopic, macroscopic and a corridor fine-tuning level, to compute a timetable explicitly driven by the above mentioned performance indicators. A case study on the Dutch railway network illustrates the feasibility of this approach to achieve the highest timetabling design level.

Keywords
Railway timetabling, Robustness, Stability, Micro-macro, energy efficiency

1 Introduction

The performance of railway operations depends highly on the quality of the timetable. In the last decade, timetabling software has become more and more common, from running time computations via mathematical timetable optimization to railway operations simulation. Nevertheless, these tools and their focus vary widely from country to country and often lack consistency since they are used independently for different purposes and do not lead to an integrated set of tools geared towards a well-defined timetable design process. In the EU FP7 project ON-TIME (Optimal Networks for Train Integration Management across Europe) one of the aims was to improve the timetabling design process with a timetabling framework that leads to improved robust and resilient timetables capable of coping with normal statistical variations and minor perturbations in operations. This paper describes the developed ON-TIME timetabling framework.

A state-of-the-art review of literature and practice revealed a lot of research in mathematical models for macroscopic timetable optimization (ON-TIME 2013), see also the review papers by Bussieck et al. (1997), Cordeau et al. (1998), Caprara et al. (2007), and Lusby et al. (2011). These macroscopic models rely implicitly on reliable input data which may not always be available. This might explain why these models and algorithms did not yet find their way into daily timetabling practice, except at the strategic level. A recent trend in the scientific literature consists of robust timetabling models (Cacchiani

and Toth 2012) that incorporate stochasticity or uncertainty in the input. Microscopic timetabling models that use a higher level of detail are limited in the literature and mainly focus on single track railways, see e.g. Brännlund et al. (1998). Also models based on blocking time theory (Hansen and Pachl, 2014) fall within this category. Most of these blocking time models are employed for computing capacity consumption using the timetable compression method or within microscopic simulation tools. Moreover, optimization models based on blocking times have been developed for real-time rescheduling, see e.g. D'Ariano et al. (2007). Recent papers apply two-level microscopic-macroscopic models to generate conflict-free timetables (Gille et al. 2008, Caimi et al. 2011, Schlechte et al. 2011). In these papers, the transformation from microscopic to macroscopic models is straightforward but the reverse is more complicated.

The timetabling practice shows a similar separation, with either macroscopic models to compute network timetables using normative input, or microscopic blocking-time based tools for detailed planning on corridors and stations but without support for network optimisation. Timetable evaluation on feasibility, stability or robustness is typically applied –if at all– after timetable construction using simulation tools with unclear procedures how the results are used to improve the timetable design. Timetabling tools are mostly concerned with routine work such as running time calculations, mostly discarding energy-efficiency, and making visualizations such as time-distance diagrams and platform occupation diagrams. Some railways (SE, UK) are starting to apply microscopic simulation tools for conflict detection as a complementary step to their macroscopic timetable planning tools. If a significant change of the timetable is foreseen either for lines or for complicated areas, robustness simulation studies are made also to ensure the feasibility of the timetable and give a rough idea of its robustness (ON-TIME 2013).

Based on the state-of-art review essential performance measures were derived that should be taken into account to achieve a good timetable (Goverde and Hansen 2013). These performance indicators include infrastructure occupation, stability, feasibility, robustness, resilience, travel times and energy efficiency. Depending on the degree that these indicators are taken into account in the timetable design process, a higher timetabling level can be obtained that lead to better timetables but at the cost of increased data requirements (Goverde and Hansen 2013).

This paper presents a new innovative three-level timetabling framework to achieve the highest timetabling level by integrating all the mentioned performance indicators in the design process. The approach is an iterative process on three levels: microscopic, macroscopic, and a corridor fine-tuning level, where each performance indicator is optimized or tested at the appropriate level. A set of implemented algorithms are described from a functional perspective as a proof of concept for this framework demonstrating the feasibility of this approach. The approach is applied to a case study from the Dutch railways showing the overall improved timetable performance.

Section 2 presents the timetable performance indicators that should be taken into account to reach a high timetabling design level. Section 3 then presents the performance-based timetabling framework with successively the functionalities of microscopic timetabling, macroscopic timetabling and corridor fine-tuning and their interactions. Section 4 illustrates the approach to a case study of the Dutch railway network and finally, Section 5 ends with conclusions and recommendations.

2 Timetable performance

The quality of a railway timetable can be measured by several Key Performance

Indicators (KPIs). Traditional KPIs are the operational speeds or scheduled running times on train lines, and more general scheduled travel times in networks including transfer times where train lines meet. On the other hand, the main KPIs of railway operations are punctuality and reliability. Short travel times in the timetable do not necessarily imply good punctuality or transfer reliability, but on the contrary they may lead to large waiting and realized travel times when connections are missed or trains cancelled. Therefore, the timetable must also be robust to normal variations of running and dwell times so that punctual and reliable operations can be realized.

Furthermore, structural route conflicts between trains due to too tight scheduling must be avoided to prevent unnecessary braking and waiting of trains with negative consequences for safety, punctuality and energy consumption. The latter point is typical for railways which are characterized by trains competing for the same infrastructure. Track capacity allocation is therefore an integrated part of railway timetable design. On this level the timetable is therefore also known as the traffic plan, which contains the exact routes of all trains and the orders of trains over conflicting routes. At this level also the safety and signalling constraints must be incorporated to prove that the traffic plan is conflict free and the infrastructure capacity consumption allows normal deviations from train paths.

The above concepts are captured in several performance measures as follows (Goverde and Hansen 2013):

- **Scheduled travel time**: The time scheduled between any origin and destination including running times, dwell times and transfer times.
- **Infrastructure occupation**: The share of time required to operate trains on a given railway infrastructure according to a given timetable pattern.
- **Timetable feasibility**: The ability of all trains to adhere to their scheduled train paths. A timetable is feasible if (i) the individual processes are realizable within their scheduled process times, and (ii) the scheduled train paths are conflict free, i.e., all trains can proceed undisturbed by other traffic.
- **Timetable stability**: The ability of a timetable to absorb initial and primary delays so that delayed trains return to their scheduled train paths without rescheduling.
- **Timetable robustness**: The ability of a timetable to withstand design errors, parameter variations, and changing operational conditions.
- **Energy consumption**: The amount of energy consumed by the train traffic.

Some of these performance measures are based on typical macroscopic quantities such as the scheduled travel times, while others require a microscopic level of detail such as infrastructure occupation, timetable feasibility and energy consumption. Timetable stability refers to a minimum amount of time allowances that must be available throughout the timetable and in particular at bottlenecks, while robustness refers to how these allowances are distributed between the train paths to maintain performance when trains deviate ‘slightly’ from their scheduled paths. Stability is closely related to infrastructure occupation and can be incorporated using the UIC guidelines on acceptable infrastructure occupation (UIC 2013) at the microscopic level, while robustness represents a trade-off with short travel times and is therefore best considered at the macroscopic level together with travel time. Energy consumption is typically a secondary objective and can therefore be considered as a fine-tuning step after the time allowances have been set based on feasibility and robustness.

3 Performance-based timetabling

3.1 Framework

The proposed timetabling approach tries to schedule all train path requests with sufficient time allowances for a stable and robust conflict-free timetable and satisfying the UIC infrastructure occupation norms (UIC 2013). This is in accordance to the Network Statements issued by the Infrastructure Managers from all EU countries to allocate the infrastructure capacity to the Railway Undertakings. This might require extending critical running times on corridors with an unacceptable capacity consumption to decrease running time differences. However, we compute timetables at a precision of 5 s instead of a minute, to avoid capacity waste and unrealizable process times by rounding to minutes.

The timetabling framework is *performance-based* in the sense that all six timetabling KPIs from Section 2 are explicitly taken into account to guide the timetable construction process. To make this possible an integrated approach is proposed on three levels:

- A *microscopic* level for highly detailed local computations;
- A *macroscopic* level for aggregated network optimisation; and
- A *fine-tuning* level for corridor optimization.

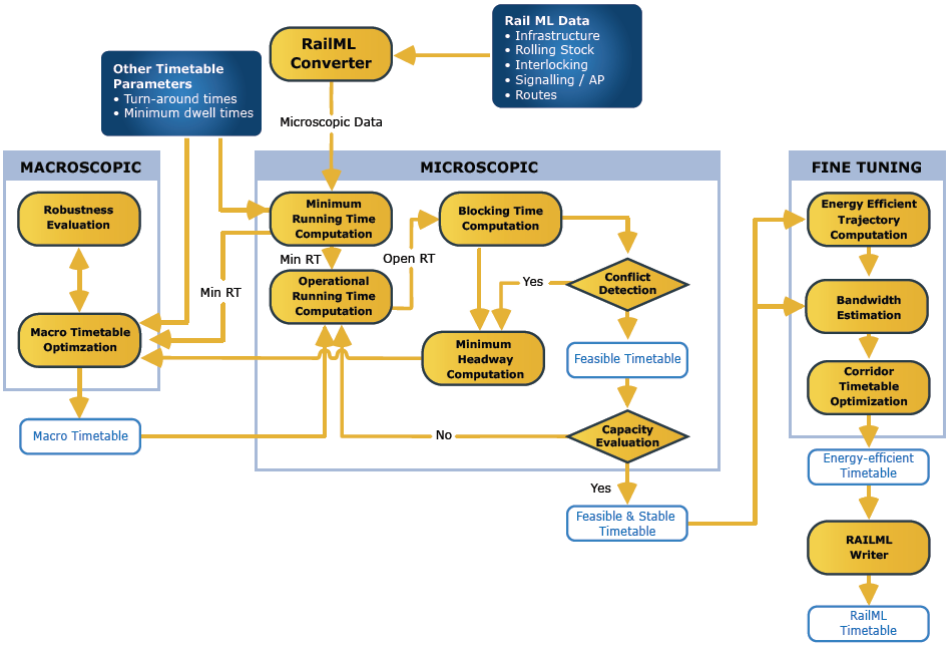

Figure 1 Three-level performance-based timetabling framework

Figure 1 illustrates this three-level timetabling approach. The input data are standardized RailML files. The microscopic model computes detailed running and blocking times, and aggregates the results into a macroscopic model that contains only the main macroscopic stations characterized by train interactions such as overtaking, connections, and merging or crossing railway lines that need decisions at the macroscopic level such as synchronization and train sequence orders. The macroscopic model then computes a network timetable taking into account network constraints and trying to avoid cancelled train path requests. The macroscopic timetable is transformed back to the

microscopic model that fills in the details on microscopic level. These two models work iteratively where the microscopic model is used for conflict detection, infrastructure occupation and stability given the (completed) macroscopic timetable, while the macroscopic model optimizes travel times and robustness given the constraints set by the microscopic model. Infrastructure occupation is based on the UIC timetable compression method (UIC 2013) which also provides norms for acceptable stability. The macroscopic model is an ILP model and includes a simulation model to find the most robust timetable out of several hundred feasible solutions. The overall cost function contains several terms including a robustness cost derived from the simulations. These micro-macro iterations converge to a timetable that is conflict-free, stable and robust (Besinovic et al. 2015).

The third level optimizes the speed profiles of all trains on each corridor between main stations while maintaining the scheduled event times at the corridor ends. In this optimization, the stochastic dwell times at the intermediate stops are taken into account and the arrival and departure times at these stops are optimized with respect to expected delays and energy savings. The input to this level is again provided by the microscopic model that computes both the aggregated data for the corridor and the bandwidths that can be used by the local trains for optimizing their speed profile. The final result is exported in RailML format extended with the scheduled speed profile information that can be used by the trains for running punctual and energy-efficiently (ON-TIME 2014a).

Algorithm 1 shows a complete list of the successive steps of the performance-based timetabling. Each of these steps is performed by a separate exchangeable module and as such the approach is general. In the remainder of the paper we will focus on the implementations carried out within the ON-TIME project from a functional point of view.

Algorithm 1 PerformanceBasedTimetabling

```
Input: railML infrastructure, rolling stock, interlocking, timetable
Output: Traffic plan at track section level in Timetable railML
Build microscopic network topology
Compute time-optimal speed profile and minimum running times
Build macroscopic network topology
Compute nominal running times by adding minimum running time supplements
Compute operational speed profiles based on nominal running times
Compute blocking times
Conflicts ← 1; Stable ← 0
repeat until Stable
  while Conflicts do
    Compute minimum local headways
    Compute macroscopic network by aggregating running times and local headways
    Compute macroscopic timetable using network timetable optimization
    Recompute operational speed profiles based on the macroscopic timetable
    Compute microscopic running and blocking times
    Conflict detection
  end while
  Compute capacity consumption
  if an unstable corridor exists
    then for each unstable corridor do
        Relax nominal and maximum running times
        Conflicts ← 1
    else Stable ← 1
  end if
end repeat
Compute energy-efficient speed profiles
Compute bandwidths for local trains
Corridor timetable optimization of local trains
Return Timetable railML
```

3.2 Microscopic timetabling

The microscopic module considers multiple functions for computing and providing necessary input to other modules as well as evaluating a timetable at the microscopic level. These functions incorporate three KPIs: infrastructure occupation, stability and feasibility. As already mentioned in Section 3.1, the module first computes the speed profiles and running times. Afterwards, the blocking times are determined which are the necessary input for conflict detection and infrastructure occupation, as well as for deriving the minimum local headway times for the macroscopic module.

The microscopic network used within the microscopic timetabling allows high detailed computations with accurate output. Arcs represent homogeneous behavioural sections defined by a constant characteristic of speed limit, gradient and radius, while the nodes present various infrastructure elements like signals, switches, stopping points or section borders. Additionally, procedures were developed for network and data transformations from the microscopic to macroscopic level, and vice versa. Details of the building blocks of the microscopic module are given in Besinovic et al. (2015a, 2015b).

In the remainder of this section we consider successively the main microscopic functionalities: speed and running time calculations, conflict detection, and infrastructure occupation and stability.

3.2.1 Speed and running time calculations

At the basis of a good timetable are well-defined running times. In particular, the scheduled running time consists of a minimum running time and an additional running time supplement. A good understanding of these two components is essential for the design of conflict-free, robust and energy-efficient timetables.

The minimum running time is the time required for driving a train from one point to another assuming conflict-free driving as fast as possible. Additionally, the corresponding speed profile represents a detailed train trajectory. The computation algorithms for speed profiles and running times have to be as detailed as possible in order to provide the high accuracy requirements. Running times are computed from microscopic train dynamics that require detailed rolling stock and infrastructure data, including route-specific static speed and height profiles. The corresponding Newton's motion equations are solved by numerical ordinary differential equation solvers (Hansen and Pachl 2014).

In regular day operations, trains are affected by stochastic variations of running and dwell times due to e.g. varying train compositions, driver behaviour, passenger volumes and weather conditions. Therefore, allowance times are added to the minimum process times so that they are robust to normal variations of the process times. These allowances must satisfy certain timetable design norms, consisting of a mix of relative and absolute values for the nominal process times (minimum process time plus minimum allowance). Running time supplements are given in percentage of minimum running time, in some countries depending on train category, while nominal dwell times are specified depending on rolling stock type and station, and nominal transfer times are provided depending on station and platform distances. The resulting nominal process times are input to the macroscopic timetable optimization as lower bounds to the scheduled process times. In the optimization the nominal times can be increased further depending on the network constraints and objective functions, resulting in the scheduled running times. The objective function of the macroscopic optimization must prevent excessive journey times by stretches of all running and dwell times. In addition an overall upper bound can be provided to the roundtrip time of trains.

Hence, in the first iteration, the minimum running times are enriched with the

minimum time supplements and as such represent the nominal running times that are used in the macroscopic model. Additional to the running time, the *operational speed profile* defines the associated train trajectory. The operational speed profile can be obtained by exploiting the available time supplements in two ways: a) cruising at speeds below the speed limits, or b) computing energy-efficient speed profiles with optimal cruising speeds and coasting. During the timetable construction with several micro-macro iterations the reduced speeds are applied as these are much faster to compute than the optimal speed profiles. In the fine-tuning these speed profiles are replaced by the energy-efficient ones.

In current practice mostly macroscopic timetabling models are used that, in a nutshell, try to assign time allowances in order to satisfy a given objective function. In this way, the running time supplements are allocated without actually testing that the resulting distribution of time supplements result in acceptable train trajectories. A big variation between two (or more) successive allocated time supplements may be problematic to reproduce a valid speed profile. Even if a speed profile is possible satisfying the given time supplements, the constructed running behaviour may be unacceptable from a practical point of view when very low cruising speeds result. For example, the German practice requires that cruising speeds may not be under 40 km/h. This may be violated in the case of a relative large running time supplement over a short section. Furthermore, it is undesirable to continuously change driver behaviour such as alternating between accelerating and decelerating with different cruising speeds. Hence, even a macroscopically feasible timetable cannot always be reconstructed at the microscopic level and consequently implemented in practice. Therefore, the operational speed profiles must be computed to test feasibility of the distribution of the time allowances.

We implemented these guidelines in the computation of operational speed profiles. The scheduled running times and corresponding operational speed profiles are computed after each macroscopic timetable computation, resulting in feasible train trajectories, which are also essential for an accurate calculation of blocking times.

The successive blocking times per train over a corridor represent a so-called *blocking time stairway*. Blocking times are computed using blocking time theory (Hansen and Pachl 2014). The blocking time of a single block section depends on the block length, the train speed, and the signalling system. It consists of a setup time, sight and reaction time, the approach time to the block section over at least the braking distance, the running time in the block, the clearing time in which the train clears the block over its entire length, and

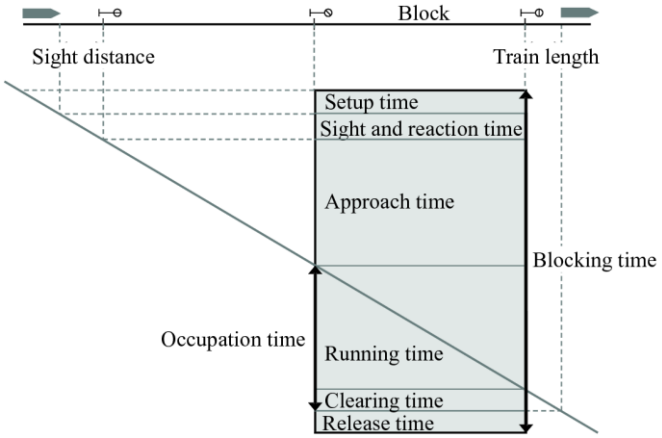

Figure 2 Blocking time of a running train

the release time of the route, see Figure 2. The blocking times are the essential input to the main microscopic algorithms such as conflict detection, capacity assessment and minimum headway computation. Recall that the blocking times are based on the nominal running times in the initial micro-macro iteration, and on the scheduled running times in all following iterations.

3.2.2 Conflict detection and realizability

Timetable feasibility is a key performance measure. It is important to have a feasible timetable in order to provide uninterrupted train runs, i.e., without unnecessary braking and re-acceleration. This timetable KPI is beneficial from several perspectives: 1) it improves safety by preventing unnecessary red signal approaches; 2) it gives less workload to drivers; 3) it provides a more comfortable ride to passengers; and 4) it saves energy. Therefore, each time a macroscopic timetable has been computed, the microscopic module automatically checks the timetable on microscopic feasibility.

The feasibility of the timetable is tested twofold: a) a *realizability* check of scheduled event times; and b) *conflict detection*. The former is simply tested by checking whether the scheduled running and dwell times exceed the minimum values. Note that the macroscopic timetabling model always provides realizable aggregated scheduled process times, so this realizability check is mainly focused on the event times at the smaller stations and other microscopic timetable points after transforming the macroscopic timetable onto the microscopic network. Unrealizable process times are mainly caused by rounding down, which becomes problematic specifically when scheduled event times must be given in minutes. In our approach, the macroscopic model computes timetables with a precision of 5 s, while we allow a precision of 1 s in the microscopic model so that rounding is not an issue anymore.

The conflict detection model determines if the scheduled trains can run undisrupted. For this blocking times are used on the basis of the operational speed profiles. Conflicts are indicated by an overlap of the blocking times of two successive trains. The second train then approaches the block section that is still blocked by the preceding train and therefore must brake in response to the signalling logic. These track conflicts are solved by shifting trains in time until their blocking times do not overlap anymore. This shift naturally initiates the change in the minimum headway between the trains. So, after all track conflicts have been detected, the corresponding minimum headways are recomputed. These new headways are given back to the macroscopic timetabling model to iteratively adjust the macroscopic timetable until all track conflicts are resolved.

3.2.3 Capacity consumption and stability

Capacity consumption is defined as the time share needed to operate trains according to a given timetable pattern taking into account scheduled running and dwell times. As such, it directly determines the stability of the timetable. The same as for conflict detection, we use the computed blocking times to evaluate the capacity consumption.

A timetable is called *stable* if any train delay can be absorbed by the time allowances in the timetable without active dispatching. Therefore, the larger the time supplements and buffer times the better is the ability of the timetable to prevent propagation of delays, i.e., the timetable is more stable. If the total amount of buffer time in a corridor is higher than the amount recommended by the UIC code 406, the timetable is considered sufficiently stable. Otherwise it is defined unstable and the macroscopic timetable has to be recomputed to reduce the infrastructure occupation on the critical corridors or stations and thereby releasing buffer times.

The recommended UIC stability norms are given in Table 1. The values presented here are for a given corridor for the peak period or the whole day. Norms for station areas still require more research as elaborated in UIC (2013).

Table 1 Recommended UIC infrastructure occupation for corridors

Type of line	Peak period	Daily period
Dedicated suburban passenger traffic	85%	70%
Dedicated high-speed	75%	60%
Mixed traffic	75%	60%

The capacity assessment model developed is based on max-plus automata (Gaubert and Mairesse 1999) and explained in Besinovic et al. (2015b). The model is applicable to both corridors and stations. For now, we assume the given UIC norms from Table 1 for both corridors and stations. If the computed infrastructure occupation is not satisfactory for a corridor then we relax the running time supplements of the trains in the corridor to allow more homogenised traffic through the considered corridor by reducing running time differences. This relaxation is explained in Besinovic et al. (2015a). The relaxed constraints are provided to a new iteration of the macroscopic timetable optimization.

3.3 Macroscopic timetabling

The macroscopic timetabling considers the railway network at an abstract level, neglecting many details of the real-world network. In particular, only timetable points like stations and junctions, where trains overtake, merge, cross or connect, as well as the lines connecting them are represented at the macroscopic level. The motivation of applying this network reduction is that it is computationally faster to work with a simplified network, and therefore several potential timetables can be evaluated according to the different key performance indicators, including robustness. Clearly, once the 'best' macroscopic timetable has been determined, its feasibility at a microscopic level is checked, as described in Section 3.1.

As explained in Section 2, the quality of a timetable is evaluated according to several performance measures. In the macroscopic model we incorporate several objectives, in order to consider these performance measures in the computation of the macroscopic timetable. They are based on the nominal running times, dwell times and connection times computed in the microscopic model, as described in Section 3.2. The macroscopic timetabling is used to determine the best feasible schedule of trains in the macroscopic network by considering a trade-off between timetable efficiency (i.e. journey times, connection times, number of scheduled trains) and robustness.

In the following, we describe the objectives and constraints that are included in the macroscopic ILP model, and present a delay propagation model that is used to achieve timetable robustness. We refer the reader to Besinovic et al. (2015a) for further details on these models.

3.3.1 Macroscopic timetable optimization

The macroscopic timetable optimization is based on the definition of a time expanded graph, built upon the macroscopic network described above: in the time expanded graph, every node corresponds to an arrival or departure of a train at or from one of the stations of the macroscopic network at a certain time of the planning horizon.

For a given train and its *route* (i.e. the sequence of macroscopic stations that the train serves or passes) its macroscopic feasible timetable corresponds to a *feasible time-distance path* in this time expanded graph that visits all the stations on the route while

respecting the given maximum *journey time* from its origin station to its destination station. This correspondence is the key element of the macroscopic ILP optimization model. More specifically, the ILP model contains a binary variable for each feasible time-distance path of any train, which specifies whether the path is selected as the timetable of the train in the solution or not. It is useful to define the variables of the model in this way, since all the constraints that are related to the feasibility of a timetable of a single train can be directly expressed through the definition of its feasible time-distance paths. However, the drawback of this model is that it contains an exponential number of variables. We can cope with this drawback by solving the ILP model in a heuristic way using a *randomized multi-start greedy heuristic* (Besinovic et al. 2015a).

We associate a *cost* to each time-distance path, which represents the quality of the corresponding timetable for the train, without taking into account the interaction with other trains. In particular, the cost takes into account the running and dwell times exceeding the nominal ones. The minimization of path costs is one of the objectives of the ILP model, which is clearly related to the performance measure of scheduled travel time described in Section 2, as it corresponds to the minimization of the journey times of the trains. The journey times in the final solution may however be larger than the feasible minimal ones, since the optimization must find a balance between minimizing running and dwell times, and the other objectives, described in the following. Train paths are scheduled by taking into account a trade-off between minimal and robust journey times.

Connection times cannot be included directly in the path cost: indeed, they refer to pairs of trains and not to single trains. However, *timetable connectivity*, i.e., the connection between pairs of trains for passenger transfers or rolling stock connections, is also taken into account as one of the objectives of the macroscopic model. In particular, we minimize the number of *missed connections*, as well as the time exceeding the *nominal connection time*. To this aim, the macroscopic model receives as input the set of train connections that should be included in the timetable, which are given as triplets (train1, train2, station) and the nominal connection time, i.e., the ideal time between the departure of train2 and the arrival of train1 at the station. We consider that a connection is missed when at least one of the two connecting trains is cancelled. Recall that we are in the planning stage of timetabling. In this stage a train cancellation corresponds to not fulfilling a request for a train service. If both trains are scheduled we compute the difference between the actual connection time and the nominal one. Both missed connections and exceeding connection times are minimized in the objective function of the ILP model. Note that similarly to what happens for journey times tight connection times may lead easily to delay propagation. Hence, tight connection times may be penalized leading to a trade-off between small and robust connection times.

Another main goal of our ILP model is to maximize the *transport volume*, i.e., the passenger or cargo-tonne delivered (ON-TIME 2014a): this is achieved in our model as the minimization of cancelled train path requests.

We consider an additional main objective: timetable *robustness*. This objective is not directly included in the ILP model or in the randomized multi-start greedy heuristic, but it is dealt with by a *delay propagation model* that will be described in the next subsection.

All the described objectives are included in the ILP model by using a weighted multi-objective function, in which different penalties are associated with the different objectives. Depending on the penalty values, one objective can have priority over another one, or the goal can be to find a trade-off between the different objectives. In summary, the multi-objective function contains the following terms, each one weighted by a penalty that is a parameter of the optimization model:

- Minimization of path costs, i.e., minimization of journey times
- Minimization of missed connections
- Minimization of time exceeding the nominal connection times, and
- Minimization of cancelled trains, i.e., maximization of the transport volume.

In order to take into account the described objectives, next to the path variables also auxiliary variables are included in the ILP model used respectively for computing the number of missed connections, the excess times over the nominal connection times and the overall number of cancelled trains (Besinovic et al. 2015a).

Timetable *feasibility* at a macroscopic level is achieved by means of the constraints included in the ILP model. In particular, feasibility is ensured by imposing that the timetable is conflict-free, i.e., it respects nominal running times, nominal dwell times, minimum headway times, and capacity constraints.

The nominal running and dwell times are respected by defining, for each train, feasible time-distance paths in the graph. In order to respect headway times and capacity constraints, we impose that at most one path, among a set of conflicting paths, can be part of the solution. Auxiliary constraints are imposed in the ILP model, in order to ensure the correct definition of the auxiliary variables for computing the objectives.

The proposed model can deal both with cyclic and non-cyclic timetabling. In the former, we are given routes for *train lines* rather than individual trains, as all trains belonging to the same line must visit the same sequence of stations. Similarly, we are given the journey time of each line and in addition the *periodicity* of the trains of the line. In order to satisfy the periodicity constraint, we impose that either all trains of the line are scheduled or all of them are cancelled. Clearly, the penalty for train cancellation is very high and therefore it is very unlikely that a complete train line will be cancelled. Different planning time horizons are to be considered for cyclic or non-cyclic timetabling. In our case study we focus on the cyclic case (see Section 4).

The ILP model is solved in a heuristic way by using a randomized multi-start greedy heuristic. This is an iterative algorithm starting each iteration from a different order of the trains. Each iteration, the trains are scheduled one at a time according to the given order. More precisely, scheduling a train corresponds to selecting one of the feasible time-distance paths of the graph for the train, i.e., fixing one of the corresponding variables in the ILP model. The choice of the best variable to be fixed is done by executing a dynamic programming procedure which takes into account all the trains already scheduled in the current iteration and therefore computes a conflict-free timetable for the current train. In addition, all the described objectives are taken into account in the dynamic programming procedure by assigning penalties to the unpromising nodes of the graph of the train, so that the best path will visit the nodes with the smallest possible penalties. In the case of periodic timetabling, at each iteration of the algorithm we select a feasible time-distance path for the entire line, i.e., we select simultaneously one path for each train of the line, therefore ensuring that the periodicity constraint is respected.

Clearly, different orders of trains can lead to different timetables as each train is scheduled by the dynamic programming procedure in the best possible way, while avoiding conflicts with the trains previously scheduled in the given order.

Once the algorithm has been executed for a given number of iterations, several macroscopic feasible timetables are available among which we would like to choose the best one. One possibility is to select the timetable with minimum cost with respect to the objectives considered in the multi-objective function. In this case, we would select a timetable with minimum journey times for the trains, minimum connection times between

connecting trains and maximum number of trains in the network. This could be a very good choice with respect to the efficiency of the railway system, but, at the operational stage trains close to each other could lead to large delay propagation as well as missed connections. A trade-off between the timetable efficiency and its robustness must be achieved to avoiding bad performance at the operational stage. As explained in Section 3.2, robustness is incorporated at a microscopic level by inserting time allowances. In the next paragraph, we explain how we consider robustness also at a macroscopic level.

3.3.2 Robustness evaluation

As previously mentioned, timetable robustness is not directly inserted in the ILP model, but computed in a post-optimization phase using a delay propagation model. This model is used to take into account the stochasticity of the events that can occur at the operational stage, such as train delays. The goal of this model is to evaluate the robustness quality of each feasible timetable determined by the randomized multi-start greedy heuristic and to select as best timetable the one having the smallest *robust cost*, given by the cost of the timetable according to the multi-objective function plus the cost of the timetable according to the delay propagation model. The latter works as follows.

A set of delay scenarios (1000 in our computational experiments) is randomly generated using a standard normal distribution. For each delay scenario the effect on each timetable is evaluated by applying a heuristic algorithm that tries to resolve the potential conflicts caused by the generated delays by retiming the trains (see Besinovic et al. 2015a for further details). The algorithm computes the overall delay propagation or establishes that some conflicts cannot be resolved. Accordingly, a cost is assigned to each timetable which takes into account the effect of all the delay scenarios on the timetable. This cost corresponds to the cost of the timetable according to the delay propagation model and is defined as the average settling time over all delay scenarios, where the settling time is the time required until all delays have been absorbed by the time allowances in the timetable.

The best timetable in terms of multi-objective value and robustness is then selected as the best macroscopic timetable and this is the outcome of the macroscopic timetabling.

3.4 Corridor fine-tuning

Energy efficiency becomes more and more important within the railway system. Currently several approaches exist for energy-efficient driving and energy-optimal conflict resolution within real-time traffic management and optimization (Hansen and Pachl 2014). However, the timetable is the static basis for real-time operation. On the one hand, the static timetable has to enable real-time operational control measures, which means that allowance times are available and provide flexibility for traffic management. On the other hand, when real-time optimization methods are applied such as energy-efficient driving, the possible real-time trajectories have to be considered already within the timetabling process in order to avoid conflicts due to the real driving behaviour. Therefore, within the ON-TIME timetabling approach energy-efficient speed profiles are already considered in the timetabling process.

3.4.1 Energy-efficient speed profiles

The energy-efficient speed profiles are computed with respect to the microscopic infrastructure and rolling stock characteristics for the given scheduled running times (including running time supplements) that were the result of the micro-macro timetabling iterations. The optimal driving trajectories are determined according to the theory of energy-efficient driving (Howlett and Pudney 1995). This trajectory is typically

characterised by different regimes and the switching points between the regimes: acceleration with maximum acceleration power, cruising at maximum speed, coasting, and braking with maximum (service) braking effort. Figure 3 shows a simplified illustration of the application of different driving regimes between two stops on a simple section with constant gradient and speed limit (Albrecht 2014).

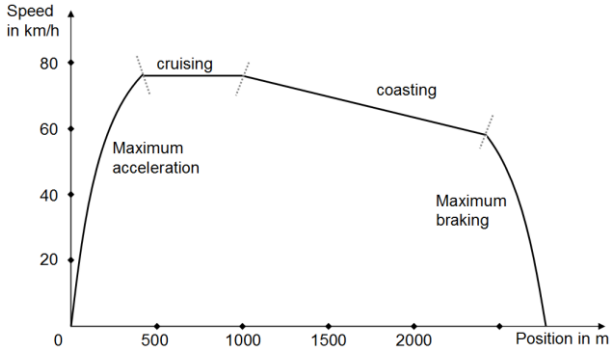

Figure 3 Energy-optimal driving regimes (Albrecht 2014)

The trajectories are used to re-define the blocking times and check conflicts within the static timetable. In addition, the information on the switching points and regimes can be used to guide optimal energy-efficient driving in case of punctual train operation even if no dynamic driver advisory systems are used. If driver advisory systems are used they essentially give dynamic speed advice with respect to delays and follow the scheduled energy-efficient speed profile otherwise.

3.4.2 Corridor optimization

The last step in the timetabling process is the corridor fine-tuning for regional trains between the macroscopic timetable points. Note that the event times at these macroscopic timetable points were optimized in the macroscopic timetable optimization. For intercity trains all served stations are important points and the energy-efficient speed profiles are already determined in the previous step. For local trains however the arrival and departure times at intermediate stops on the corridors were not yet optimized and they offer flexibility for optimization within given time windows, see Figure 4.

The bandwidths are determined from the blocking times of the trains preceding and following the local train that has to be optimized. Hence, the trajectories of the neighbouring trains are important in order to maintain a conflict-free timetable. The total amount of running time supplement over the corridor and the bandwidth between the macroscopic timetable points are provided by the macroscopic and microscopic timetable levels, respectively. Given these parameters, the published arrival and departure times at the intermediate stops for the local trains can still be optimized. In this optimization the stochastic dwell times are the most influencing factors.

On the one hand, the published times are important for passenger arrivals and delay calculations in case of long dwell times. On the other hand, these published times are restrictive because early departures are not allowed in case of small dwell times. During timetabling the dwell time allowances could be exchanged with running time allowances where they could be applied for e.g. energy-efficient driving in case of short dwell times.

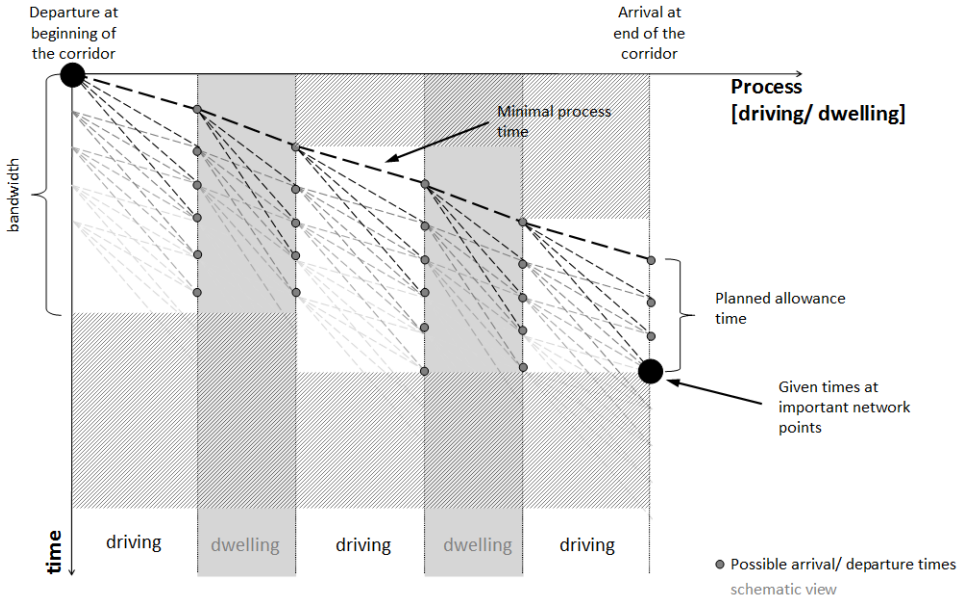

Figure 4 Flexibility of the corridor optimization

Figure 5 explains the dependency between the dwell time distribution, the departure of the train and the corresponding energy consumption. The figure at the top shows that a short planned dwell time leads to a possible punctual departure of the train and less energy consumption because the allowance time could be used for additional running time. If the dwell time is a slightly higher this leads to a little delayed departure and higher energy consumption on the following section. In contrast to this the bottom figure shows the pessimistic published departure time (for a higher predicted dwell time). In this case, the probability for a delayed departure is less, but the probability of waiting for the departure time is higher. Therefore, the minimal achievable energy consumption is higher than when publishing an earlier departure time. This means that dwell times should not be considered as deterministic in the timetabling process but as dwell time distributions within the process of finding the published departure times. The dwell time distributions must correspond to the realized dwell times and must be obtained using operational data. This enlarges the robustness of the timetables for the local trains.

The target of the corridor optimization is consequently to determine the published arrival and departure times at intermediate stops within the given bandwidths under consideration of the stochastic dwell times and enabling energy efficient driving in case of short dwell times. The mathematical approach is another two-stage optimization process in order to find the optimal timetable on a corridor where the objective function is a weighted sum of energy consumption and expected delays at the intermediate and final station (Binder and Albrecht 2013).

Finally, the energy-efficient speed profiles provide additional information to train drivers. Particularly, the detailed train trajectory computed by the microscopic and fine-tuning module could be adopted as static driver information as well as input to traffic management systems that need a trajectory at track section or signal level.

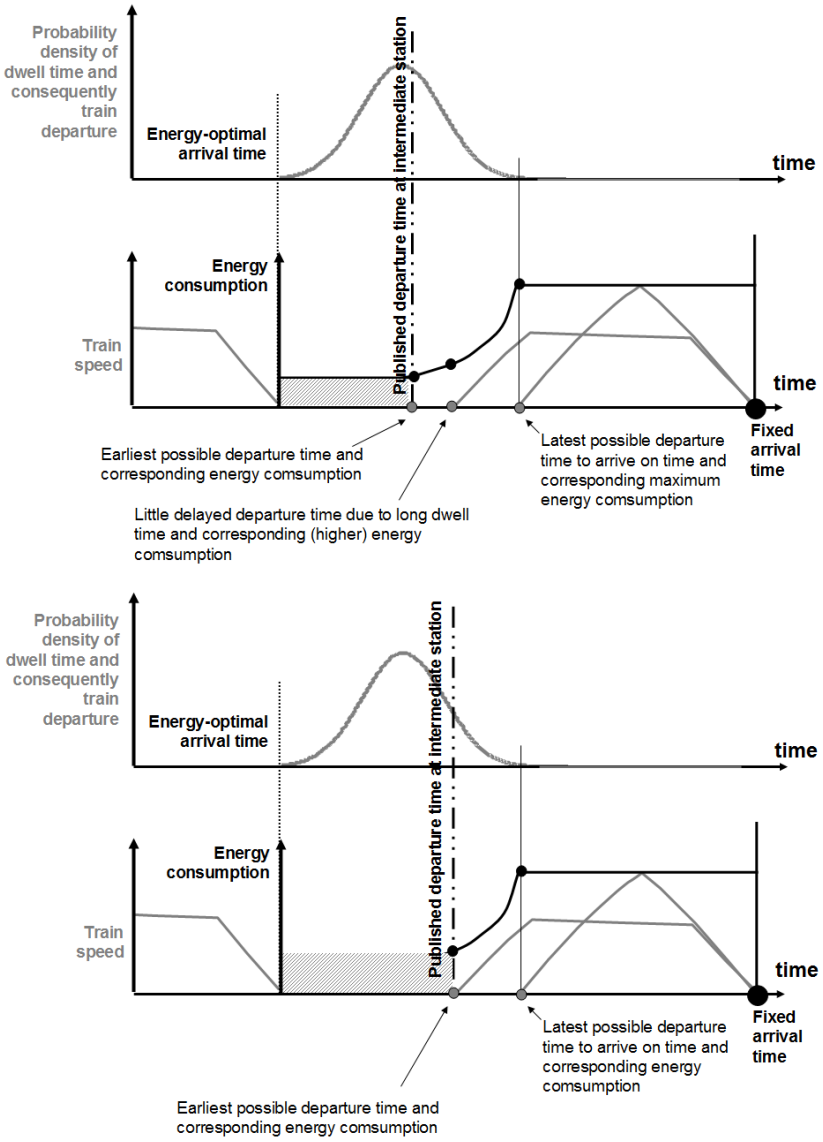

Figure 5 Dependency of published departure time and energy consumption

4 Case study

The performance-based timetabling approach has been applied on a case study of a central part of the railway network in the Netherland (ON-TIME 2014b), consisting of the railway network bounded by the four main stations Utrecht (Ut), Eindhoven (Ehv), Tilburg (Tb) and Nijmegen (Nm), with a fifth main station ‘s-Hertogenbosch (Ht) in the middle and 20 additional smaller stations and stops. Four corridors connect Ht to the other main stations. The train line plan in this part of the network is taken from the 2011

timetable and consists of four intercity lines and six local train lines with a frequency of two trains per hour each, see Figure 6. The intercity lines 800 and 3500 offer a regular 15 min service between Ut and Ehv but have different origin/destinations outside this area. The regional line 13600 from Tb to Ht continues as the line 16000 from Ht to Ut, and vice versa. The line 9600 from Ehv couples in Ht to the line 4400 to Nm, and vice versa. In addition, an hourly freight path with maximum speed of 120 km/h is scheduled from Ut-Ehv. So overall, 41 trains are running per hour in this network. As an illustration of our results we focus on the corridor Utrecht-Eindhoven in this paper.

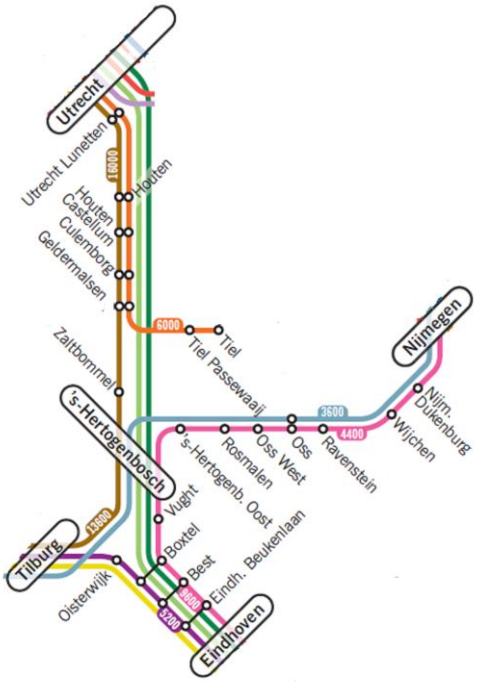

Figure 6 Passenger line plan of the Dutch case study

Figure 7 shows a time-distance diagram of the computed hourly timetable for the corridor Ut-Ehv. The vertical axis shows time in minutes downwards. The horizontal axis shows distance with the station positions indicated. The blue lines are IC trains, the magenta lines are local trains, and the green line is the freight train. Note that the sections Btl-Ehv and Htn-Htnc have four tracks. Figure 8 shows the corresponding blocking time diagram for the route of intercity train line 3500. Note that only the blocking times are shown for all trains running on the same tracks as train line 3500. The gaps in the blocking time stairways for some trains correspond to running on parallel tracks in stations or the four-track lines between Htn-Htnc and Btl-Ehv. Around Ht also some blocking times are visible corresponding to crossing trains from/to Tilburg or Nijmegen.

The optimized timetable shows periodic passenger trains with regular 15 min services of both IC and local trains where two similar train lines follow the same route. Hence, effectively 15 min train services are realized instead of two separate 30 min train lines. The ICs overtake the local trains at Geldermalsen (Gdm) in the southbound direction, but not in the return direction. The fast freight train departs after the local train from Utrecht Centraal (Ut) and overtakes this local train at the four-track line around Houten (Htn).

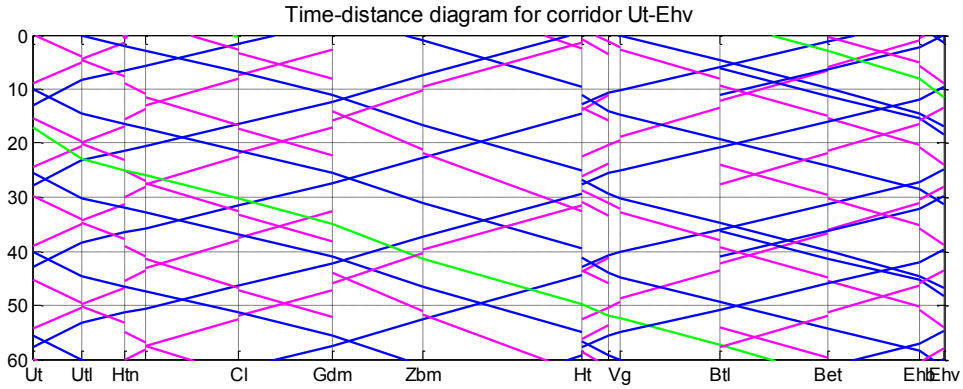

Figure 7 Time-distance diagram corridor Utrecht – Eindhoven

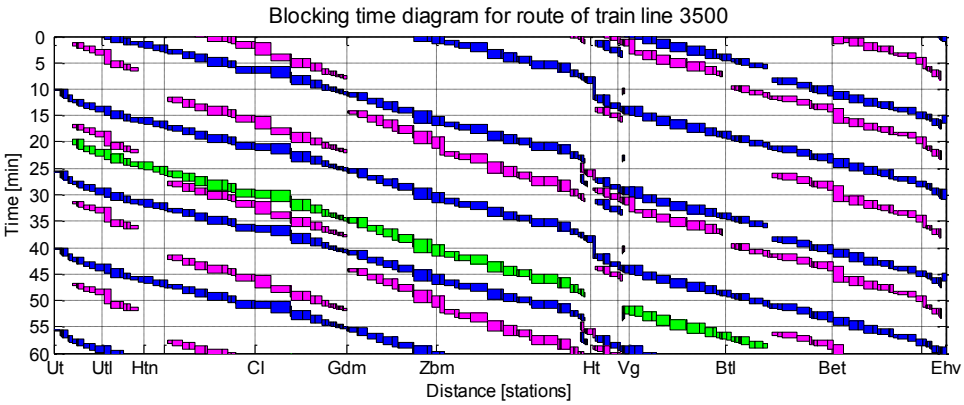

Figure 8 Blocking time diagram corridor Utrecht – Eindhoven

The blocking time diagram of Figure 8 shows no overlapping blocking times and hence illustrates that the timetable is conflict-free. Moreover, the timetable is robust illustrated by the buffer times (white space) between the train paths. Only between Houten Castellum (the station just after Htn) and Culemborg (Cl) the freight path and the next local train are tight so that a slight delay of the freight train might propagate to the local train but the buffer time between this local train and the next IC prevents further knock-on delays. In Gdm, the local train also has a longer dwell time that can be used to recover from an arrival delay. In the absence of the freight train the situation is robust, which is the usual case currently with on average one freight path per two hours on this corridor.

Table 2 gives the infrastructure occupation of the main corridors and stations, respectively. All the infrastructure occupation percentages are below the recommended stability value of 60% defined by the UIC for mixed traffic corridors in daily periods, which was one of the constraints of the timetabling algorithms. Corridor Ut-Ht is the heaviest used one with infrastructure occupation 57.8%. Ht has the highest infrastructure occupation of 58.3%, which includes also the crossing routes from/to Tilburg and Nijmegen. The relative low infrastructure occupation of corridors Ht-Ehv and back is due to the four tracks between Btl and Ehv.

Table 2 Infrastructure occupation

Corridors			Stations		
Corridor	**Time [min]**	**Ratio [%]**	**Station**	**Time [min]**	**Ratio [%]**
Ut-Ht	34.7	57.8	Btl	15.7	26.2
Ht-Ut	32.1	53.4	Ehv	15.7	26.1
Ehv-Ht	22.0	36.7	Gdm	15.7	29.5
Ht-Ehv	24.2	40.3	Ht	35.0	58.3
			Htn	15.0	25.0
			Ut	20.9	34.8
			Vga	17.2	28.7

Table 3 Journey times

O-D	**Minimum journey time** [min]	**Scheduled journey time** [min]	**Journey time increase** [%]
Ut-Ehv	44.9	48.2	7.3
Ehv-Ut	47.6	51.3	7.8

Table 4 Energy consumption all trains

Speed profile	**Energy consumption** [kWh]	**Energy saving** [%]
Minimal-Time	64 395	-
Reduced cruising speed	58 800	8.7
Energy-optimal	41 667	35.3

Table 3 gives the average journey times over all trains running over the complete corridor from Ut to Ehv or backwards in a basic hour, i.e., eight IC trains and one non-stop freight train of 120 km/h speed limit. The minimum journey time refers to the minimum running and dwell times while the scheduled journey time includes the time supplements. On average, the time allowances over the complete corridor are 7.3% and 7.7% for the southbound and northbound directions, respectively, which can be exploited for energy-efficient driving.

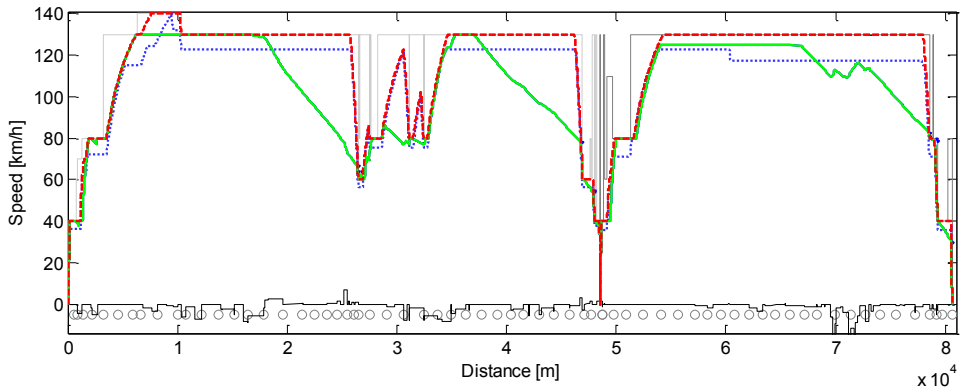

Figure 9 Speed profiles: static speed limit (solid grey), time-optimal (dashed red), reduced cruising speed (dotted blue), and energy-optimal (solid green)

Figure 9 illustrates the various speed profiles for the intercity line 3500 Ut-Ehv with intermediate stop in Ht. The bottom of the figure indicates the gradients (solid black line) and the signals (grey circles) over the line. The dashed red line is the time-optimal speed profile corresponding to the minimum running times, while the dotted blue line is the operational speed profile with the running time supplements distributed over the line using reduced cruising speeds. The solid green line is the energy-optimal speed profile with

clear coasting regimes before the areas with speed restrictions. Table 4 gives the total energy consumption of all trains running in the network of the case study, so 21 trains with all passenger trains counted once (corresponding to a basic half hour timetable including the freight train). With respect to the minimum running times the running time supplement saves 8.7% energy consumption when cruising at a reduced speed and even 35.3% using the energy-optimal speed profile with coasting. As was illustrated in Figure 9 for the IC 3500, the time supplements of the trains are distributed well over the corridor so that coasting could be applied very effectively.

5 Conclusions

This paper presented a performance-based timetabling approach and illustrated it to a case study from the Netherlands showing good results on all performance indicators. In particular, the approach highlighted eight recommendations that need to be considered explicitly in the design of a stable robust conflict-free timetable with optimal journey times:

- Microscopic calculations of running and blocking times taking into account all running route details at section level (gradients, speed restrictions, signalling)
- Microscopic conflict detection guaranteeing a conflict-free timetable
- Timetable precision below 15 s to minimize capacity waste
- Incorporation of (UIC) infrastructure occupation and stability norms
- Macroscopic network optimization with respect to journey times, transfer times, cancelled train path requests and associated cancelled connections
- Macroscopic robustness analysis using stochastic simulation to obtain the most robust network timetable
- Stochastic optimization of timetables for local trains on corridors taking into account stochastic dwell times at intermediate stops
- Energy-efficient speed profiles computed and incorporated for all trains.

Moreover, standardized exchange files such as railML for infrastructure, rolling stock, and the timetable is recommended, where the presented timetabling approach generates an output Timetable railML with scheduled train paths at (track-free detection) section level, extended with scheduled energy-efficient speed profiles.

Acknowledgement

The research leading to this paper was funded by the European Union's Seventh Framework Programme (FP7/2007-2013) in the ON-TIME project under Grant Agreement SCP1-GA-2011-285243.

References

T. Albrecht, 2014. Energy-efficient train operation. In: I.A. Hansen, J. Pachl, 2014 (Eds.). Railway Timetabling & Operations. Eurailpress, Hamburg, pp. 91-116.

N. Besinovic, R. Roberti, E. Quaglietta, V. Cacchiani, P. Toth, R.M.P. Goverde, 2015a. Micro-macro approach to robust timetabling. Paper presented at RailTokyo 2015.

N. Besinovic, E. Quaglietta, R.M.P. Goverde, 2015b. Microscopic computer-aided tools for automated railway traffic planning. *Proceedings of the 95th Annual Meeting of the Transportation Research Board*, Washington DC, 11-14 January, 2015.

A. Binder, T. Albrecht, 2013. Timetable Evaluation and Optimization under Consideration of the Stochastical Influence of the Dwell Times. *Proceedings of the 3rd International Conference on Models and Technologies for Intelligent Transportation Systems 2013*, pp. 471-481.

U. Brännlund, P.O. Lindberg, A. Nou, J.E. Nilsson, 1998. Railway timetabling using Lagrangian relaxation. *Transportation Science*, 32(4), 358-369.

M.R. Bussieck, T. Winter, U.T. Zimmermann, 1997. Discrete optimization in public rail transport. *Mathematical Programming*, 79, 415-444.

V. Cacchiani, P. Toth, 2012. Nominal and robust train timetabling problems. *European Journal of Operational Research,* 219(3), 727-737.

G. Caimi, F. Chudak, M. Fuchsberger, M. Laumanns, R. Zenklusen, 2011. A new resource-constrained multicommodity flow model for conflict-free train routing and scheduling. *Transportation Science*, 45(2), 212-227.

A. Caprara, L. Kroon, M. Monaci, M. Peeters, P. Toth, 2007. Passenger railway optimization. In C. Barnhart, G. Laporte (Eds.), *Transportation*, Handbooks in Operations Research and Management Science, 14, Elsevier, Amsterdam, pp. 129-187.

J.F. Cordeau , P.Toth, D. Vigo, 1998. A survey of optimization models for train routing and scheduling. *Transportation Science*, 32(4), 380-404.

A. D'Ariano, M. Pranzo, I.A. Hansen, I.A., 2007. Conflict resolution and train speed co-ordination for solving real-time timetable perturbations. *IEEE Transactions on Intelligent Transportation Systems*, 8(2), 208–222.

S. Gaubert, J. Mairesse, 1999, Modeling and Analysis of Timed Petri Nets using Heaps of Pieces, *IEEE Transactions on Automatic Control*, 44(4), 683-697.

A. Gille, M. Klemenz, T. Siefer, 2008. Applying multiscaling analysis to detect capacity resources in railway networks. *Computers in Railways XI*, WIT Press, Southampton, pp. 595-604.

R.M.P. Goverde, I.A. Hansen, 2013. Performance Indicators for Railway Timetables. In: *IEEE International Conference on Intelligent Rail Transportation (ICIRT)*, Beijing, pp. 301-306.

I.A. Hansen, J. Pachl, 2014 (Eds.). Railway Timetabling & Operations. Eurailpress, Hamburg.

P. Howlett, P.J. Pudney, 1995. *Energy-Efficient Train Control*. Springer, London.

R. Lusby, J. Larsen, M. Ehrgott, D. Ryan, 2011. Railway track allocation: models and methods. *OR Spectrum*, 33(4), 843-883.

ON-TIME, 2013. *Assessment of State-of-Art of Train Timetabling*. Report ONT-WP03-I-EPF-008-03.

ON-TIME, 2014a. *Methods and algorithms for the development of robust and resilient timetables*. Report ONT-WP03-D-TUT-034-01.

ON-TIME, 2014b. *Benchmark analysis, test and integration of timetable tools*. Report ONT-WP03-D-TUT-037-02, Deliverable 3.2.

T. Schlechte, R. Borndörfer, B. Erol, T. Graffagnino, E. Swarat, 2011. Micro-macro transformation of railway networks. *Journal of Rail Transport Planning and Management*, 1(1), 38-48.

UIC, 2013. *Code 406: Capacity*. International Union of Railways, Paris, 2nd Edition.