State-of-the-art of railway operations research

I. A. Hansen
Delft University of Technology

Abstract

The paper discusses the current state of research concerning the modelling of stochastic train operations in order to optimise the use of capacity by means of analytical methods and micro-simulation. Timetable quality is governed by precise running and realistic recovery times, as well as optimal headway and buffer times. Analytic (queueing) models and micro-simulation are used for estimation of waiting times, while combinatorial models and stability analysis by means of max-plus algebra technique are suitable for network timetable optimisation. Operations quality can be modeled by the estimation of (admitted) queue length and micro-simulation. A recently developed probabilistic model for the estimation of delay propagation in stations and junctions based on the stochastic variation of track occupancy times depending on conditional probability distributions of hinder by headway constraints and route conflicts is presented, while stochastic timetable optimisation may help to determine a more efficient distribution of running time margins and to increase punctuality in large networks.

Keywords: timetabling, waiting time, delays, rescheduling, level of service.

1 Introduction

So far, railway timetables are based on deterministic running, dwell and headway times between stations. These times are mostly scaled in minutes and refer to a virtual stopping point at the stations. Small variations of the service times are compensated by standard running time and dwell time supplements, as well as margins between the train paths. The amount of time supplements and margins applied for scheduling, however, is mainly based on rules of thumb, sometimes checked by simulation, and only seldom derived from statistical analysis of real-world operations data.
Stochastic analytical approaches, like queueing models [1–5] assume the train intervals and the service times to be independent random variables, which is questionable in case of periodic timetables with high frequencies. The variation of inter-arrival times and minimal headway times is modelled mostly on the basis of assumed random distributions for inter-arrival and service times. The estimated waiting time of a timetable is a function of the track occupancy and the coefficients of variation of the scheduled headway and service times. Waiting times generated by the stochastic variables of the timetable are clearly distinguished from estimated knock-on delays during operations.

Optimisation models based on linear programming [6], in general, apply deterministic variables for searching the optimal value of an objective function, like minimisation of overall running times in large complex networks, at given constraints e.g. minimal headway and transfer times between trains. The first known stochastic optimisation approach for estimating the robustness of railway timetables was recently presented by Vromans [7].

Micro-simulation models, like STRESI [8], RailSys [9], OpenTrack [10] or ATTPS [11] are used to estimate the effect of exogenous random primary delays on track occupation and consecutive delays of hindered trains. The induced primary delays are drawn from assumed or empirical distributions for a given timetable, track infrastructure and signal system. Kaminski [12] introduced a heuristic limit for the buffer time distribution at bottlenecks in order to compensate for 80% of the primary delays. Carey and Carville [13] presented a heuristic approach for solving conflicts between train paths and routes and simulation of random delays in order to test the reliability and robustness of timetable options.

This paper will discuss the influence of basic stochastic variables for timetable design and its impact on railway performance, as running and recover time, minimal time headway and punctuality. A new approach for the estimation of blocking times and buffer times at bottlenecks by means of blocking time ‘waves’ is presented. The paper concludes with the remaining issues for further research and development.

2 Timetable quality

The evaluation of timetable quality must reflect the requirements of the different stakeholders, as customers (passengers, shippers), infrastructure managers, train operating companies and public authorities (government, passenger transport executives). The timetable itself represents the transport supply, which should comply best with the market demand. There exist numerous criteria to evaluate the quality of railway services, which are more or less important for customers, infrastructure provider, train operators and public authorities. The quality of timetable design for a given traffic demand, rail infrastructure and train mix depend mainly on:

- Precise and realistic running times
- Sufficient but not too large recovery times
• Exact minimal headway times between different pairs of trains,
• Estimated waiting time.

2.1 Running and recovery times

Stochastic train running times were first modelled by Steckel [14]. He performed multiple simulations (400) with random input for different locomotive hauled trains and found Weibull distributed running times independent from the route and train characteristics with positive skew and parameter $\delta$ between 1.2 and 3.0. The peak factor (ratio of pure running time plus margin/ pure running time) varied strongly due to route conditions and train mass (Fig. 1). An approximate estimate of the superior value based on simple regression functions of the share of acceleration time is used to calculate the necessary running time margin of a certain route without intensive simulations.

![Figure 1: Frequency of running time peak factors of passenger trains (source: Steckel).](image)

Empirical distributions of exact running times can be easily derived from track occupation and release data which are recorded and saved automatically by the existing signalling and safety system. As these data only contain the passing times per train number at main signals and these are distant from the stopping points at platforms the additional deceleration and acceleration time respectively from and to the last (first) main signal needs to be estimated taking into account the remaining distance between the last/first measuring point and the stop location, as well as the train characteristics (length, weight, power, standard deceleration/acceleration rate). At TU Delft the tools called TNV-Prepare and TNV-Filter [15] have been developed by which the recorded passing times at signals and insulation joints of any train can be filtered and the real running times and delays per train series be computed automatically according to their route, type and period of the day.

The statistical analysis of the precise running times per train line enables to determine the percentile of running times within a certain level of probability (Fig. 2). Furthermore, the true minimal running times per link can be revealed at
a precision of seconds by selecting only those trains that were delayed at the preceding departure station and experienced no hinder.

![Figure 2: Distributions of running speed of a train series in case of (no) hinder North of the station The Hague HS (September 1999).](image)

As the running times depend not only on the alignment, number of and dwell times at intermediate stops, type, length and weight of rolling stock operated, but also on the behaviour of the train drivers and the weather conditions, the specific performance distributions can be sorted and used as information for quality management and fed back to timetable designers and operations personnel.

The running time distributions per train line, thus, can be used for specifying the amount of recovery time to be applied in timetables instead of using standard supplements for all links and trains. The recovery times needed for achieving a certain punctuality level of the train arrivals based on departure delay distributions can be estimated numerically by means of convoluting the integrals of the departure delays and running times per link and train line [16].

### 2.2 Minimal headway and buffer times

In practice, timetable designers mostly apply standard mean minimal headway times (2 to 5 min) between train paths depending on the type of conflict and train sequence. The existing buffer times in a conventional graphical timetable which indicate only the train paths and headway times at a scale of minutes cannot be determined sufficiently, because these do not indicate the precise start and end of the capacity consumption according to the prevailing signalling an safety system.

The minimal time headway between two trains on a railway route is governed by blocking times of trains and its speed differences [17, 18]. The blocking time graph of a train represents the time instances that a train needs to run safely without hinder at design speed over a sequence of track sections. The scheduled train paths virtual as well as the track occupancy times are represented in a time-distance diagram and indicate clearly the remaining buffer times. Any overlap
between the blocking times of different trains clearly indicates a timetable conflict, which needs interaction in order to avoid a deceleration or stop of the following train.

The blocking times depend not only from the signal spacing and train length, but also from the actual train speed and deceleration rate. If the movement authority for the train at sight distance of the distant signal is given late because of insufficient headway to the preceding train, the following train would be decelerated automatically, which means an increase of the blocking time. From empirical analysis of train detection data at different Dutch railway stations follows that the mean speed of trains approaching to a station is about 10 to 20 % less than the design speed and may, in case of hinder, even be 50% less [19].

Furthermore, the scheduled dwell times are often exceeded during operations, which lead to an increase of the blocking time of the routes serving the platform tracks. The quality of timetable design would be enhanced considerably if the estimated blocking times reflect well the variation of train speed and dwell times in practical operations.

So far, the blocking times in timetables are assumed to be deterministic and estimated with a precision of seconds. The arrival and departure times of trains and the headway times between trains in most railway timetables, however, are determined currently with a precision of minutes due to rounding-up of the estimated running times and easy comprehension by the passengers. This practice includes hidden scheduled waiting times, which could be exploited if the determination of the arrival, departure and headway times (for in-company planning and operation purposes) was done at a higher precision in steps of 5 to 10 sec (according to current practice in Japanese railways).

Figure 3: Division of an interlocking arrangement into route nodes [18].

For capacity estimation of heavily occupied routes in bottlenecks the track infrastructure needs to be subdivided in individual route nodes: the smallest track elements that can be used by one train at a time (Fig. 3). The variations of time headway and train speed at heavily occupied junctions may lead to knock-on delays and queueing of trains before network bottlenecks. On the routes approaching to level crossings close to main stations we observed a significant drop of the mean train speed by a bout 20 to 50% with regard to design speed.
which means a clear reduction of infrastructure performance. This means the scheduled speed and headway times must correspond better to real operations and the train drivers need to be supported by dynamic speed advices in order to avoid hinder to and by other trains at junctions. In fact, the blocking times are stochastic, and not deterministic, and so are the remaining buffer times (Fig. 4).

![Figure 4: Distribution of blocking times of a train series at the station The Hague in case of (no) hinder (September 1999).](image)

The probability density of the capacity consumption of the trains passing at bottlenecks could be represented by stochastic blocking time ‘waves’. The range of blocking times covered by the stochastic waves of each train line indicates the extra use of capacity due to the variation of real operations in comparison with the deterministic blocking time estimated on the basis of the scheduled running time and possibly dwell time per train.

The total track occupancy and the remaining timetable slack of a given line is estimated finally by means of a virtual compression of the blocking time graphs according to the new UIC norm 406 capacity [20]. It recommends maximal track occupancy rates of 75% during peak hours and 60% a day for lines used in mixed traffic.

### 2.3 Waiting times

Any timetable comprises scheduled waiting times and generates, in practical operations, non-scheduled waiting times. Scheduled waiting times result from differences between the scheduled and desired travel, headway, departure or arrival times. The latter depend on the market demand, track possession times due to maintenance and conflicts between resulting different train graphs at bottlenecks, whereas non-scheduled waiting times can emerge from technical failures of track infrastructure or rolling stock, accidents, and train delays. Scheduled waiting times may occur randomly and its amount is used as indicator
of timetable quality, whereas non-scheduled waiting times are experienced as train delays during operations. The latter serve as indicator of operations quality.

The higher the reliability of tracks and rolling stock, the higher is their availability for operations and the practical capacity. The total waiting time during a certain time period, in general, increases exponentially with the number of trains operated. The challenge is to find an optimal track occupancy that enables a desired level of service, which can be defined by a certain amount of waiting time.

Schwanhäusser [1] developed already in 1974 a stochastic approach for the estimation of the mean queue length as function of the distribution of primary delays, buffer time, mean headway, train sequence and priority by a model of the type $M/D/1$. This led e.g. to an estimated mean buffer time of 1 min in case of a minimal headway of 2 min and all trains delayed (p. 59). As the assumed Poisson distribution of train arrivals near bottlenecks could not be confirmed later on, other queueing models of the type $G/G/1$ and $M/G/1$ have been developed by Wakob [2] and Hertel [3] respectively.

Hertel developed further an analytical approach for the determination of the optimal train intensity, which is situated between the minimal relative timetable sensitivity (partial derivative of mean waiting time to track occupancy) and the maximal traffic ‘energy’ (defined as product of train intensity and speed). The range of optimal use of track capacity as function of the waiting time, speed and ‘energy’ of the route, in general, lies at about 150 to 200 trains per day, while the mean waiting time per train would be up to about 10 min.

Wendler [4] extended in 1999 the existing two-train queueing model to a three-train model of type $G/G/1$ in order to estimate the impact of available time lags between the headway times on the main track of a junction for intermediate merging and crossing of trains. This approach is suitable for modelling the process of independent random requests for infrastructure capacity from different train operators and estimating the resulting waiting times due to timetable adaptation of train paths, headways and synchronization.

The relative errors of the estimated waiting times by the different queueing models with regard to the waiting times computed by micro-simulation in the case of the railway node Mannheim/Ludwigshafen as a function of track occupancy for exponential inter-arrival times are shown in Fig. 5. The 2-train queueing model of Wakob (thin line) generates exponentially increasing errors for individual route nodes at higher occupancy rates of 0.7, whereas the Light-Traffic approximation of the 3-train model (bold line) of Wendler performs best up to 0.5 occupancy and the Heavy-Traffic approximation is the best analytical estimation for occupancy rates between 0.6 and 0.8.

Kaminski [12] analysed the blocking and buffer times in the existing timetable, as well as the recorded train delays for a large network of more than 2000 km in Germany and found a good fit of the buffer times with a negative-exponential distribution. The available buffer time was less than 1 min (standard minimal buffer time of Deutsche Bahn) in about 8% of the train sequences. He investigated the impact the length of buffer times that correspond to 80% of the recorded train delays at a number of stations by means of multiple simulation
while maintaining the train order. The results showed a mean consecutive delay per delayed train of about 30 sec and a reduction of 9% of the estimated hinder between trains. Thus, the expected operations quality could be improved by means of tailored buffer times per route section as function of the distribution of train delays.

![Figure 5: Relative error of waiting time estimation by different queueing models compared to micro-simulation (source: Wendler).](image)

Microscopic simulation tools like STRESI [8], RailSys [9] OpenTrack [10] respectively estimate the total waiting times of lines and networks by means of computing the impact of multiple drawing of random primary delays from a predefined distribution on total waiting times. The virtual hinder generated by modification of the scheduled blocking time graphs is transformed automatically into suitable conflict resolution measures (bending of train paths, increasing of dwell times, change of train order and routing), while the impact is estimated by the distribution of the total amount of delays, the kind and percentage of hindered trains and the punctuality level.

The timetable slack of interconnected lines in networks with periodic timetable depends not only on the buffer times of the individual lines, but also on the buffer times between the scheduled train arrivals and departures at the transfer stations. The following constraints for the design of an integrated network timetable hold:

- The round trip time of a line must be integer multiple of the headway,
- The travel time between the nodes (sum of the running and dwell times) must be an integer multiple of half of the cycle time (mostly 60 min),
- The scheduled arrival and departure times of interconnected lines at the nodes must overlap sufficiently,
- The sum of the travel times on a circuit between three nodes must be an integer multiple of the cycle time and of the time headway.

As these constraints lead to a great complexity of periodic timetables in large networks with many nodes and lines, combinatorial methods by means of mathematical programming have been developed to find an optimal solution for integrated network timetables [6]. The network timetable problem is known as NP-complete and requires the development of intelligent algorithms to solve
within a reasonable computation time – if a solution for the given constraints exists – and/or to relax some timetable constraints in order to achieve an approximate solution. The predefined minimal headway and transfer time constraints applied in combinatorial models don’t guarantee its practical feasibility in complex junctions and need to be proven by a detailed analysis of blocking times.

The existing buffer times and its distribution in periodic timetables of large networks can be calculated analytically by means of the (max, +) algebra approach, which is characterised by a transformation of the timetable constraints into a set of simple recursive linear equations [21]. This approach enables to identify the critical circuits within the network and to calculate the minimal cycle time of the timetable, as well as the existing timetable slack. A tool for the automatic transformation of conventional periodic network timetable constraints into a (max,+) state matrix of the travel times, minimal headway times and transfer times, for critical circuit and recovery time analysis has been developed. The impact of an increase or decrease of travel times and buffer times on the timetable slack and on the location of the critical circuit of periodic network timetable can be estimated rapidly, as well as the propagation of train delays in the network.

3 Operations quality

The operations quality can be assessed by means of different analytical approaches and micro-simulation. A sufficient level of performance requires, of course, that the basic timetable is feasible and the scheduled running times, dwell times and headways are realistic, which means the timetable must not contain any conflict between train paths proven by the absence of overlaps between the blocking time graphs. The impact of primary delays on the robustness of train services is expressed e.g. by the queue length, amount of knock-on delays, the number of involved trains, links and stations, as well as the fading-out time. Given a certain distribution of primary delays, the consecutive delays may not grow steadily due to insufficient buffer times. The scheduled times and track occupancy, thus, must allow for recovery.

3.1 Queueing model of RWTH Aachen

Schwanhäusser [1] defined the smoothness of operation as the share of unscheduled waiting trains at any location in order to proceed. He developed an analytical approach for estimating the consecutive delays in case of hinder based on queueing theory. The admitted level of hinder during operations was defined empirically by a maximal queue length (equal the maximum total consecutive delays per day), which is given by a simple exponential equation as function of the share of passenger trains:

$$L_W = \alpha \cdot e^{-\beta \cdot p^P}$$

$L_W$ : queue length [-]

$\alpha, \beta$ : coefficients
Three different levels of quality of operations (0.5=very good, 1.0=satisfactory, 1.5=unsatisfactory) are defined by the ratio of the estimated length of queue divided by the maximum admitted length of queue. The standard maximum amount of knock-on delays to be expected due to operation of a certain number of trains admitted by Deutsche Bahn is 130 min up to 300 min per day depending on the share of passenger trains. These threshold values of maximum waiting time, however, look somewhat arbitrary and seem to need further empirical validation. The estimation of the scheduled and unscheduled waiting times as a measure of operations quality of a given timetable, infrastructure and signalling system is implemented in the software tool ANKE developed at RWTH Aachen [22].

3.2 New probabilistic model of TU Delft

The weakness of queueing models in case of periodic timetables can be overcome by an advanced probabilistic model of the process times of each train line (running or arrival time, dwell time, departure time) as stochastic variables, while the hinder due to signalling and route conflicts by delayed trains and its priority are taken into account [16].

The approach is based on the stochastic variations of track occupancy times e.g. the variations of train speed in case of different signal aspects with conditional probability distributions. Distinction is made between

- hindrance of approaching and departing train respectively by late release of preceding block or route section and
- Hindrance of approaching and departing train respectively by late release of crossing (merging) of routes.

The model is solved on the basis of a numerical approximation of the Stieltjes convolution of individual independent distributions.

The delay propagation to a train approaching to a station is modelled depending on the actual signal aspects, which may force the train to decelerate or even to stop. The arrival time distribution can be estimated by the probability that the train experiences more or less hinder and the conditional probability of the arrival time in case of each condition. This distribution is formulated as,

\[ F_{A_i^p}(x) = P[A_i^p \leq x] = \sum_{h=1}^{3} P((A_i^p \leq x) | (\text{Condition } h)) \cdot P[\text{Condition } h] \]

The departure of a train at station is modelled analogously as function of the probability of dwell time extension and route conflicts and the distribution of running times until release of the routes. For more details of the analytical model see [16]. The calculation of the convolution is done numerically by applying the composite trapezoidal rule. The proposed model has been validated by means of train detection data recorded at The Hague HS station. To validate the model, we obtained the required input distributions on the basis of empirical track occupation and release data.
3.3 Stochastic timetable optimisation model of EU Rotterdam

Vromans [7] developed recently a timetabling model based on a Periodic Event Scheduling Problem (PESP) approach and combined with a reliability evaluation by means of simulation with a fixed order of events. The delay propagation is calculated by performing a number of consecutive runs of the scheduled trains, while these are perturbed by exogenous random disturbances and measuring the resulting arrival delays. The objective of the linear programming model is to minimise the mean arrival delay over all runs by means of an optimal allocation of the running time supplements along the line. If the standard 7% running time supplement of the Dutch Railways was distributed optimally instead of linearly, the punctuality of the trains operated in the investigated cross-country corridor Haarlem-Maastricht would be increased significantly.

3.4 Micro-simulation

The propagation of non-scheduled waiting times (delays) in large networks of interconnected lines can be modelled also by micro-simulation. Watson [23] compared the characteristics of current commercial simulation tools and concluded that ‘signal berth’ level tools are necessary for stochastic simulation. Simulation tools mostly require interaction of the user in case of conflicts between blocking time graphs or the application of a predefined automatic conflict resolution strategy. For the evaluation of different dispatching measurements on the stability margin and the location of network bottlenecks in case of disturbance different options can be computed. The analysis of individual link and train dependent recovery times would allow a variety of experiments to estimate the robustness of different re-scheduling options (re-timing, re-ordering, re-routing). The effectiveness of different conflict resolution strategies can be evaluated on the basis of the amount of disturbance (primary delay, knock-on delays, punctuality) and fading-out time.

4 Conclusions

The current methods and tools for railway timetable design enable a high precision of the estimated travel times, headway times and time margins in order to achieve high-quality conflict-free deterministic timetables for lines and networks. Queueing models and micro-simulation tools have been applied successfully to improve timetable quality and to estimate stochastic train delay propagation on lines, in stations and networks. The key for high-quality timetabling is a precise estimation of blocking times based on realistic running, dwell and headway times taking into account the signal spacing and train processing at critical route nodes and platform tracks. Statistical analysis of empirical track occupation and release data in Dutch stations has revealed that the trains mostly leave later and operate at speeds less than scheduled due to drivers’ and conductors’ behaviour, lack of movement authority and routes still locked for other delayed trains. Deterministic models for estimation of the stability or optimisation of network timetables, like max-plus algebra technique.
or linear programming, may not represent real operations and lead to suboptimal solutions.

As queueing and simulation models for estimation of unscheduled delays in daily traffic, too, do not reflect sufficiently the impact of speed variations and behaviour of railway staff and may be based on invalid input distributions, new stochastic analytical approaches have been developed recently in the Netherlands. The probabilistic model of TU Delft enables an estimation of the survival rate of knock-on delays at platform tracks and junctions based on the distributions of the involved train pairs. It allows an estimation of the admitted maximal number of trains and distribution of consecutive delays at a given primary delay and level of punctuality. The stochastic timetable optimisation model of EU Rotterdam helps to find an optimal distribution of running time supplements along a line that may increase punctuality significantly.

Acknowledgements

This publication is a result of the research project ‘Modelling and analysis of operations in railway networks: the influence of stochasticity’ supported by the Dutch research foundation STW and Transport Research Centre of Delft University of Technology.

References


