A polynomial function of primary delays and cumulative delays on railway lines

Fabrizio Cerreto[1*], Otto Anker Nielsen[1], Steven Harrod[1].

[1] Department of Management Engineering, University of Denmark, Lyngby, Denmark

[*] Corresponding author: facer@dtu.dk, +45 45256548. ORCID 0000-0001-8392-5743.

ABSTRACT

Railway planners have several tools to simulate the operation. Simulation allows the evaluation of the railway system's quality by different points of view. Analyses of timetable stability and robustness include several indicators: recovery time and total delay among others. This paper formulates a model to describe the delay on a railway line. The total delay is characterized as a polynomial function of the primary delay given to a train.

We define a simplified delay propagation model to derive the individual delay of a generic train at a station. This is the combination of the hindrance by the previous train and the delay of the same train at the previous station, reduced by means of scheduled buffer times and timetable allowance, respectively. The individual delay is described as a linear function of the individual train number and station position. The total delay results from the summation of the individual delays and results in a cubic function of the primary delay, for small values of primary delays.

The model makes it possible to compute the total delay on a railway line with very limited use of micro-simulation, with yet high accuracy. Microsimulation is used only for the initial model calibration. Previous studies proved the validity of the model on suburban railway lines.

This is significant for the future development of models to analyze timetable stability and robustness. As an industrial application, the model will allow fast and accurate analyses of the timetables, and on-line decision support for rescheduling or dispatching.

1. INTRODUCTION

One of the most important factors for passengers' perception of railway transport is the reliability and punctuality of the service, strictly connected to operation stability and robustness. In order to measure and evaluate this, several indices and evaluation methods have been proposed in scientific research (f. i. (Vromans, Dekker, & Kroon, 2006), (Mattsson, 2007), (Salido, Barber, & Ingolotti, 2008)). The present paper reveals these methods, and presents a new closed-form analytic expression that can be used for fast analyses in the initial strategic planning phases of railway operations.

Many techniques to evaluate railway stability and robustness have been proposed in the last fifteen years, some of them based on analytical or stochastic approaches, others based on empirical models, simulation, and what-if analysis, and yet others based on ex-post statistics of the realized operation. Every technique is characterized by an inherent trade-off between performance, accuracy, and flexibility, that makes it the best suitable a particular phase of the planning and timetabling process. The knowledge on the timetables and operating conditions is typically low in the strategic phases, and increases during the process. At the same time, the analyses can be fitted with increasing details, making them more accurate, but also heavier and slower to perform. This is in contraposition with the actual needs and availabilities in the different phases. Strategic phases usually require fewer details than the tactical ones, but the latter suffer from a smaller availability of computation time. The real time operation is the hardest phase to address, as the maximum level of details is required together with extreme rapidity to support the decisions of the dispatchers live.

Stability appraisal methods include Vroman, Dekker, & Kroon (2006): they propose the relation between timetable homogeneity and the propagation of delays due to interdependencies between trains, while the inequality of headway time dispersion is investigated by Carey (1999). Salido, Barber, & Ingolotti (2008) assess the robustness in terms of average delay per train and settling time, to evaluate different solutions of their rescheduling model.

Mattsson (2007) uses a microsimulation tool to study the interferences between trains under different capacity utilization values: he finds this to be the most precise way to analyze secondary delays, but it needed very detailed input and is a time consuming process. The detailed input and output give flexibility to the method, making it adaptable to different contexts. Signaling systems, operation rules, rolling stocks, and infrastructure alignments can be modelled through the micro-simulation of any guided transport system. Pellegrini, Marlière, & Rodriguez (2014) found limitations using such a detailed method in real time applications: they concluded that the simulation of speed variation dynamics is too hard to compute in re-scheduling problems, compared to the accuracy gain. Nevertheless, it is effective for estimation of actual train interactions through the signaling system: the robustness increment given by better acceleration and braking performance can be investigated though micro-simulation to compare different types of rolling stock. Cerreto (2015) investigates the variability of some stability indicators under different capacity utilization, settling time, total delay and average delay per train, among others. The total delay results as a clear polynomial function of the primary delay that should be further investigated, and microsimulation is employed in the study, resulting an appropriate tool to assess the relation between primary and secondary delays.

Some theoretical models have been developed to support the findings of the experimental studies on the delay propagation. Among these, Hasegawa et al. (1981) use a hydrodynamic analogy to model the train traffic: they modelled the delay propagation as a shock wave in a compressible fluid and appraise the total delay as a cubic function of the primary delays by means of propagative velocity. Landex (2008) elaborates a delay propagation model computing the transfer of delay between trains through the scheduled buffer times. This model is used to study the relation between the capacity consumption and the development of the disruptions, but does not take into account the recovery of train delays through the timetable allowance. Goverde (2010) represents the timetables in timed event graphs to model the delay propagation on railway networks. The models uses a macroscopic algorithm based on max-plus algebra, therefore does only considers the interaction between trains at the stations, neglecting the interferences through the distancing

system on the open line. Stochastic models have also been introduced to predict the total expected delay from the passenger point of view: Kirchoff (2014) studies the delay accumulation and propagation on the networks. He uses an analytical formulation to derive the passengers' delay distribution, given the some primary delay distributions on the network. Meester and Muns (2007) derive the distributions of secondary delays from the primary delay distributions using phase-type distribution, to formulate a model between a general queuing model and a very detailed simulation model. Their model is suitable for larger networks, but suffers from assumption of independent running times of the trains, so only the variation of the process times is actually taken into account, while more detailed information is needed on the actual interactions between the trains.

All the authors that approached timetable analyses through micro-simulation agree on it being time-consuming, both in the set-up phase and in the running phase. These characteristics make it difficult to implement this tool in the operational phases, that require the fastest solutions. On the other hand, this is the phase that would benefit the most from an increase of the accuracy. Micro-simulation is not implemented in real-time decision support systems for railway dispatchers because it is too heavy for live computation. The smart use of micro-simulation models would lead to a reduction of computational times keeping high level of detail, making it possible to implement real-time computation. This would also apply to stochastic analyses, which need many simulation runs to achieve a required reliability. Microscopic and macroscopic simulation models, can be combined with analytical models to compute the detailed effects of given disruptions in advance, and fine tune the results with fast computation on the need.

The present paper provides a close formulation of the total delay as a polynomial function of the primary delay, under a set of assumptions. The relation is derived from a theoretical formulation of the delay propagation between consecutive trains. We computed, in first instance, the individual delay of trains at each station of a railway line, as a function of the primary delay assigned to a train at a station. Then, the individual delays of the train sum up to the total delay on the line. Finally, we validated the model using the micro-simulation of a suburban railway line in Denmark.

The importance and contribution of this finding is the possibility to fit simulation models with a close-formed analytic expression of delays. This operation reduces the costs and the processing time, and makes it possible to proceed faster and evaluate more alternatives in the initial planning phases of railway projects. The formulation can also be used for backward engineering, e.g. to assess what level of timetable supplements is needed to reach the desired reliability of the service. It also opens new paths to assess the timetable stability and robustness, through the estimation of the average buffer and supplement times on the line by mean of the micro-simulation.

1.1. Reading guideline and notation

The paper is organized as follows: an introduction on railway operation and capacity is provided in section 2. Railway simulation models are presented and classified in section 3. A delay propagation model is proposed in section 4, followed by the mathematical formulation of the individual and total delay as functions of the primary delay. A micro-simulation model of a suburban railway line north of Copenhagen was used to validate the model, and the results are presented alongside a numerical example in section 5. A discussion is presented, finally, in section 6.

2. RAILWAY OPERATION AND CAPACITY

Traffic density and occurrence of disturbances in railway operation are often positive correlated (Wiklund, 2002), and the extent of disruptions is also strongly affected by the traffic intensity relative to capacity consumption. The planned operation (published timetable) needs some buffer times to be capable to absorb disturbances. A timetable and railway system that can cope with disruptions and unexpected events without significant modifications to the operation is referred to as robust (Takeuchi & Tomii, 2005). Some of the tools to gain stability in timetables are the running time supplements and the buffer times in headways between trains; the reduction of heterogeneity in trains' characteristics is also indicated as a method to increase robustness (Salido et al., 2008).

Stability of railway operation is defined as the ability of a plan to withstand unexpected events that generate delays. It can be described by different comparisons between the magnitude of the disruptive events and the disturbance that they generate. For this reason, it is necessary to introduce a classification for delays. In the literature (Goverde & Hansen, 2013), train delays are usually classified into *primary* and *secondary delays*.

The *primary delays* are unexpected extensions of the planned times of the individual processes included in the scheduled. Rolling stock and infrastructure failures can be the causes of primary delay, as well as large passenger flows at the platform that require longer dwelling times than scheduled. Causes can also be operations related, e.g. a train driver arriving too late after a break.

The *secondary delays*, on the other hand, are delays generated by operation conflicts, which are themselves due to primary delays. When a train is delayed, it needs to use infrastructure elements like switches, crossovers, track sections, and platforms at different times than planned. A conflict arises when two or more trains request to use the same element at the same time: they will be queued by dispatching decisions, since each item can only be used by one train at a time. The delay that is generated because of the queuing is called secondary delay. The secondary delays include the delays that propagate from one train to the following ones on railway lines, because each blocking section of the line can only be occupied by one train at a time. This effect is called propagation of delays and is the subject of several scientific studies ((Goverde, 2010), (Hasegawa et al., 1981), (Meester & Muns, 2007)).

The *stability* of a timetable can be estimated by comparing some given primary delays with their effect on the whole operation (Cerreto, 2015). The total delay is the sum of the delays of all the trains on a line and it is used as an overall measure of the disruption. The timetable stability is described by the comparison between the amount of primary delays and the total delay generated because of the propagation and other conflicts in operation. The timetable *robustness* is the ability to withstand variations in design variables, and changes in operational conditions, with a limited number of variations in the plan. It is meant to minimize secondary delays in perturbed operation, and can be studied by means of the time needed to

absorb given disruptions (Goverde & Hansen, 2013). This time is called recovery time or settling time (Salido et al., 2008).

The connection between primary and total delays is a key characteristic of a timetable, which planners consider when creating the schedules. This is why methods to estimate the total delay are important for railway timetabling. The study of a functional relation between primary and total delays on railway lines is also relevant for stability and robustness analyses.

3. SIMULATION MODELS FOR RAILWAYS

Simulation models are used to simulate real processes in a computer environment in order to understand and predict the phenomenon in focus. Rail traffic simulation represents the complex railway system with the interaction between trains, infrastructure, schedules, and complex operational rules. Train movements are constrained by the infrastructure since trains can only overtake each other at designated points in the network, which usually require one of the two trains to stop, waiting to be passed. Such a highly constrained operation generates the phenomenon of delay propagation, due to the peculiar interaction between different trains, and between the infrastructure and the trains. Railway simulation models can thus be used to understand and predict the railway operation. The models can be classified by scale, approach and process (Siefer, 2008).

Scale-wise, a simulator can be classified into microscopic, mesoscopic, or macroscopic, according to the amount of detail included in the model. Macroscopic models only include aggregate information about the topology of the network, representing the stations as nodes and the lines as edges in between. Only few details are incorporated, including the minimum headways between trains at stations and fixed running time on the edges. Microscopic models, on the other hand, represent the railway system in much more detail, including the topology of each track section represented, together with the signaling system and the blocking sections. The train running times are computed by means of integration of the motion equations of the trains, given the tractive effort, the resistances and the instructions from the signaling system at every

instant of time. Mesoscopic models are a compromise between Micro- and Macroscopic models, and contain selected details, like the tracks schemes without the single blocking sections. Macroscopic models are lean, fast and not detailed, whereas microscopic models usually require hard computation and give detailed results. This makes the first suitable for large network simulation, while the latter fit better with the simulation of short lines or stations (Radtke, 2008).

Approach-wise, simulation models can be stochastic or deterministic, according to the analysis to perform. Deterministic simulation is used to study the timetables as they are or with determined operational conditions. A conflict-free timetable is the result of a deterministic design with no stochasticity in process times. Stochastic models include delay distributions, variations in process times, and they are used to design stable timetables.

Process-wise, the simulation models can be classified into Asynchronous and Synchronous. The difference is the processing order of train paths. An asynchronous model typically elaborates entire train paths and computes the interactions between trains after a priority order. Trains that are elaborated first have the highest priority, and they are not affected by trains that are elaborated later. Synchronous models, on the contrary, elaborate all the train paths at the same time, following time steps at which each path is computed according to the operation conditions.

The model introduced in this paper gives an estimation method of the total delay on a railway line as a function of the primary delay given to a train. According to the classification given above, the model can be characterized as macroscopic, deterministic, and asynchronous. It has no detail on the tracks layout and only contains information about the buffer times between trains and the scheduled running time supplements. The total delay is quantified, given a deterministic primary delay. The train paths are considered consecutively. The first trains are modelled first, so the following train cannot influence the previous ones.

4. MODEL DERIVATION

The purpose of the model development is to derive a close-form analytical expression that explains the total delay on a railway line, as a function of the primary delay of a train. The total delay is computed as the sum of the delays of individual trains at each station, which is the result of the propagation of the delay from the previous train to the following ones. The development of the delay of a train throughout the line and between consecutive trains is studied by means of buffer times and timetable allowance. The summation of the individual delays is developed under different hypothesis in this paragraph and two different polynomial relations are presented between the primary delays and the total delay. For an easier comprehension, we enclose here below a table of abbreviations to summarize the terms used in the mathematical formulation.

Table 1 - Table of abbreviations

d	Total delay measured on the line
i	Train index
s	Station index
$d_{i,s}$	Individual delay of train *i* at station *s*
t_d	Delay threshold
$a_{i,s}$	Running time supplement of train *i* between stations *s-1* and *s*
a	Running time supplement of every train between any pair of stations
$b_{i,s}$	Buffer time at station *s* between trains *i-1* and *i*
b	Buffer time at any station between any pair trains
p	Primary delay
n_s	Number of stations of a railway line
v	Traffic volume, number of trains in a timetable
s^*_i	Recovery station for train *i*
i^*_s	First train on-time at station *s*

1.2. Total delay and individual delay

The total delay *d* represents the magnitude of the disruption impact on the line: it is the total deviation from the scheduled timetable and can be calculated as the sum of every train's delay at all the stations

$$d = \sum_{i,s}(d_{i,s} \mid d_{i,s} \geq t_d), \qquad (1)$$

with $d_{i,s}$ being the delay of train i registered at station or timing point s (difference between real and scheduled time), and t_d being the delay threshold, under which we consider the train punctual. For a simpler formulation, the delay threshold will be omitted in the rest of the paper.

The delay of a train at one station is the result of the propagation of the delay from previous trains and the recovery of its own delay from the previous segment of line. The timetables include two different types of slack to reduce the delay propagation between trains and to increase each train's ability to recover delays along their paths, both of which have impact on punctuality, capacity and total travel time. The *timetable allowance*, also called running time supplement, is the difference between the minimum running time of a train on a given railway segment and its scheduled time. This supplementary time can be used to recover trains from delays. The size and distribution of the supplement times along the train paths often follows the guidelines from the International Union of Railways (UIC, 2000) or by the national Railway Infrastructure Managers rule-of-the-thumb guidelines. A large amount of supplement time improves the overall punctuality of trains, but it also increases the scheduled travel times, reducing the attractiveness and the efficiency of the railway system. Different strategies for the supplement allocation are also influential on the punctuality, and some of them are analyzed by Schittenhelm (2011), and by Siefer and Fangrat (2012).

Another element to cope with delays is the *buffer time* between trains, defined as the difference between the minimum time headway between trains at a given timing point and the scheduled interval. This prevents small delays from spreading to the following trains and increases the punctuality as well, but it also consumes capacity, as it decreases the number of trains that can be run in a given time interval (Goverde & Hansen, 2013).

1.3. Notation and problem formulation

The notation to describe the propagation of delays among trains is described in the following. $a_{i,s}$ is the timetable allowance of train *i* between stations *s-1* and *s*. $b_{i,s}$ is the buffer time at station *s* between trains *i-1* and *i*. These introduced variables are useful to describe how the delay is propagating to consecutive trains and how every train recovers over the stations.

As mentioned before, the individual train delay $d_{i,s}$ can be either the result of hindrance from previous trains or a residual delay that has been recovered only partially after its generation. In the first case, the delay of the previous train surpasses the buffer time and the exceeding part is transferred to the following train. In the latter case, the delay of a train at a station exceeds its achievable recovery. Both the cases can be modelled in the following equation, where $d_{i,s-1}$ is the delay of the same train at previous stations and $d_{i-1,s}$ is the delay of previous trains at the same measuring station:

$$d_{i,s} = max\{(d_{i,s-1} - a_{i,s}), (d_{i-1,s} - b_{i,s}), 0\} \qquad (2)$$

$a_{i,s}$ represents the timetable allowance scheduled for train *i* from station *s-1* to station *s*; $b_{i,s}$ is the buffer time scheduled between trains *i-1* and *i* at station *s*.

We formulate the individual delay as the maximum between the train's own residual delay and the hindrance from the previous train because we prevent the generation of further primary delays, so only the largest component is transferred.

The total delay on the line is the sum of the individual train delays over the stations. The individual train delays, formulated as above, are non-linear functions of previous delays. This is because the timetable allowance and buffer time change between pairs of stations and pairs of trains, so the larger component of the composite function described in (2) is different for any combination of trains and stations. To linearize the function, we introduce two assumptions, so the formulation of the individual delay can be generalized for any train *i* at any station *s*.

The model assumes that the timetable allowance is equally spread over the train paths and that all the paths are identical. It is also assumed that the buffer times between trains are equal between any given pair of

train at every point along the line. With these two assumptions, both the timetable allowance and the buffer time can be written as train- and station-independent, and the indices *i* and *s* can be omitted. Therefore, *a* is the timetable allowance of every train between *any* pair of stations and *b* is the buffer time at every point between *any* pair of trains. These simplified assumptions are reasonable in suburban railway lines, if the trains run with the same stopping pattern, and if the train paths are repeated identically over time at a given frequency.

We define *p* the primary delay of the first train at the first station, that means $d_{1,1}=p$. This delay will propagate to the following trains through the expression of individual delay given in (2). The expression can be rewritten as follows, based on the primary delay and the unique values of *a* and *b*:

$$d_{i,s} = p - (s-1)a - (i-1)b, \qquad (3)$$

Subject to the non-negativity constraint $d_{i,s} \geq 0 \; \forall \; i, s$.

This is a linear function of the individual secondary delays of trains, as a function of the primary delay, the timetable allowance and the buffer time. Given a fixed primary delay, the function is monotonically decreasing through the trains and the stations, since $a,b \geq 0$ and $i,s > 0$. The individual delay of every train at every station is thus maximum as large as the previous train or station.

It is worth to notice the dimensionality of the variables used: *a* and *b* are times, as well as $d_{i,s}$ and *p*; *s* and *i* are dimensionless quantities.

For a better comprehension, the table below summarizes the modelled individual train delay at the stations. The table includes a finite line with a given number of stations n_s and a finite number of train, namely the traffic volume *v*.

In the following two situations, the total delay is modelled as the summation of the individual train delays under the separate hypotheses of no recovery or total recovery of the delay on the line. The two situations differ with regard to the domain of summation of the individual elements.

Table 2 - Individual train delay at each station under the hypothesis of equal running time supplements and buffer times.

Station / **Train**	**1**	**2**	**…**	**s**	**…**	**n_s**
1	p	p-a	…	p-(s-1)a	…	p-(n_s -1)a
2	p-b	p-a-b	…	p-(s-1)a-b	…	p-(n_s -1)a-b
…	…	…	…	…	…	…
i	p-(i-1)b	p-a-(i-1)b	…	*p-(s-1)a-(i-1)b*	…	p-(n_s -1)a-(i-1)b
…	…	…	…	…	…	…
v	p-(v-1)b	p-a-(v-1)b	…	p-(s-1)a-(v-1)b	…	p-(n_s -1)a-(v-1)b

1.4. The total delay without recovery

Table 2 shows a linear relation between the primary delay p and any non-negative individual train delay. If the primary delay is large enough, no train will completely recover its delay before the last station n_s is reached. The total delay under this assumption can be computed summing the individual train delays, with the summation extended to all the trains and all the stations:

$$d = \sum_{i,s} d_{i,s} = \sum_{i=1}^{v} \sum_{s=1}^{n_s} (p - (s-1)a - (i-1)b) = \\ = p\, n_s\, v + \frac{1}{2}(\, n_s\, v\, (a+b) - a\, v\, {n_s}^2 - b\, n_s\, v^2). \tag{4}$$

This equation expresses the total delay as a linear function of the primary delay, given the number of trains and stations, the timetable allowance and the buffer time.

1.5. The total delay with recovery

The formulation becomes more complicated when the trains are able to recover from their delay before the end of the line. As opposite to the previous case, the primary delay is small enough to allow all the trains

to recover completely before the end of the line. It is also so small that it does not propagate to the last train at the first station, where the primary delay occurs.

The summation domain must be, therefore, limited to the trains and stations with non-negative delay, as follows:

$$d = \sum_{i,s|d_{i,s}>0} d_{i,s} = \sum_{i,s|d_{i,s}>0} (p - (s-1)a - (i-1)b)\,. \qquad (5)$$

We mentioned before that the individual delay $d_{i,s}$ is monotonically decreasing over the trains and the stations. For this reason, the domain of trains and stations with non-negative delay is a convex area, and it is still possible to split the summation in partial summations over trains and station.

We can approach the summation of the individual delays first over the trains and then over the stations, or in the reverse order. In the first approach, the summation of the individual delays over the trains returns the total delay at each station. The grand total delay is the sum of the total delays at stations. The opposite approach sums first the total delay of each individual train on the line. In this case, the grand total delay results from the summation of the individual train total delays.

The boundaries of the partial summations are calculated by means of the *recovery station* and the *first train on time*, which we define hereunder.

The *recovery station* of each train is the first station where the train has recovered its delay completely. For each train *i*, the recovery station s^*_i can be found by rounding up to the next unit the result of the generic equation

$$s^*_i = s|d_{i,s} \geq 0 = \frac{p-(i-1)b}{a} + 1. \qquad (6)$$

Due to the individual delay's monotonicity, every train is subject to a smaller delay than the previous train, and the first train has the furthest recovery station. We calculate this station from equation (6), with *i=1*:

$$s^*_1 = \frac{p}{a} + 1, \qquad (7)$$

Similarly, for any given station *s*, the *first train on time* i^*_s can be calculated by the generic equation below

$$i^*_s = i|d_{i,s} \geq 0 = \frac{p-(s-1)a}{b} + 1 \qquad (8)$$

and rounding up the result to the next unit.

The first station is where the primary delay occurs, and where the highest number of trains experience secondary delays. The first train running on time at this location is expressed by equation (8), with *s=1*.

$$i^*_1 = \frac{p}{b} + 1. \qquad (9)$$

These values set the new boundary of the summation to find the total delay: the individual delay should only sum trains at stations which individual delay is not negative. As explained above, the grand total delay can be computed either summing the total delay at individual stations or summing the total delay of individual trains. The two approaches described above will provide the same result. We give hereunder only the formulation of the first approach, which first takes into account the delayed trains at every station, and then sums the total delays at every station. The summations should only be extended to the last delayed train at each station, and to the last station where the first train is delayed with non-negative delay, i.e. $(i^*_s - 1) = \frac{p-(s-1)a}{b}$ and $(s^*_1 - 1) = \frac{p}{a}$.

Equation (5) becomes, then,

$$d = \sum_{i,s|d_{i,s}>0} d_{i,s} = \sum_{s=1}^{\frac{p}{a}} \sum_{i=1}^{\frac{p-(s-1)a}{b}} p - (s-1)a - (i-1)b = \\ = \frac{(a^2+3ab)}{12ab}p + \frac{a+b}{ab}p^2 + \frac{1}{6ab}p^3 \qquad (10)$$

The result is a polynomial relation of third degree between the primary delay and the total delay. The total delay can hence be expressed as a cubic parabola function of the primary delay. Contrary to the previous formulation, the total delay in the case with recovery does not depend either on the number of stations on the line or on the traffic volume. This characteristic is intuitively understandable: the delay does not propagate to the last train, so additional trains would not be included in the summation of the total delay;

similarly, all the trains recover the delay before the last station, so additional stations would not be included in the summation.

This formulation introduces an approximation given by the discrete nature of the quantities. Both the upper limits of the summations are given by a ratio and should be rounded up to the unit. Such a rounding cannot be computed in a double summation, so the summation is solved considering the upper limits as continuous, assuming that the error is small enough. Landex (2008) overcomes the issue including the rounding up in the summation. He only considers the buffer times between trains, that allows to include only a single summation in the formulation and solve it analytically. Our error can be computed numerically comparing the actual summation of the individual delays and the modelled summation. We assume an estimation error small enough to be neglected, and push its estimation to future research.

Intermediate cases between the two presented bounds are also possible. In particular, the combination of primary delay, timetable allowance and buffer time can be so that the first train is able to recover before the last station, but the delay is anyway propagated to the last train at the first station. Similarly, a situation can occur in which the first train is not able to recover before the last station, but the delay at the first station is not transmitted to the last train, so that there is still at list one train that is not delayed at all.

1.6. The total delay in the general case

The two cases above describe the total delay in the boundary cases of no recovery or total recovery before the end of the line. They can also be seen as two different sections of the same curve; if we fix the buffer time and the timetable allowance, the total delay as a function of the primary delay becomes a composite curve that builds up from polynomials of different degree. A third degree function describes the first section of the curve, while a linear function characterizes the last segment. The borders of the first section are the null primary delay and such a value corresponding to either the first train not being able to recover completely, or the last train being delayed at the first station. The more restrictive value is given by the ratios between the timetable allowance and the buffer time, and between the number of stations and the traffic volume.

The conditions mentioned above can be written as

$$(s^*_1 - 1) \leq n_s \rightarrow p \leq a\, n_s, \tag{11}$$

$$(i^*_1 - 1) \leq v \rightarrow p \leq b\, v. \tag{12}$$

Therefore, the total delay is a cubic function of the primary delay in the interval $p \in [0, \min(a\, n_s, b\, v)]$. On the other hand, the total delay becomes a linear function of the primary delay when no train is able to recover the delay before the end of the line. This condition is formalized by the inequality

$$d_{v,n_s} = p - (n_s - 1)a - (v - 1)b > 0, \tag{13}$$

where $d_{v,ns}$ is the individual delay of the last train $i=v$ at the last station $s=n_s$. The inequality is respected for values of primary delay greater than

$$p > (n_s - 1)a + (v - 1)b, \tag{14}$$

It is reasonable to state that within the segment between the cubic and the linear sections there is a sub-segment with a quadratic dependency between the total delay and the primary delay. In the first segment, the upper bounds of both the summations are linear function of the primary delay p, through the first train on time i^*_s and the recovery station s^*_i, so the resulting total delay is a third degree polynomial function of the primary delay. In the linear segment, both the upper bounds are fixed values, corresponding to the number of trains v and the number of stations n_s, so the resulting total delay is a linear function of the primary delay. In the section in between, either of the two upper bounds is fixed and the other still a linear function of the primary delay, so the resulting total delay is a second-degree polynomial function of the primary delay.

5. CASE STUDIES

In this section, we present numerical examples of the cases formulated above and the simulation of a real suburban railway in Denmark to show the relation between the primary delay given to a train and the total delay generated on the line.

5.1. Numerical examples on a short corridor

A case without recovery

A timetable with $v=6$ trains is operated on a railway line with $n_s=8$ stations. The timetable includes the same supplementary time between all the stations $a=2$ minutes, while the buffer time between consecutive trains is $b=6$ minutes everywhere. With primary delays over $p=44$ minutes (14), all the individual delays of every train at each station of the line will be non-negative, as shown in the table below. The total delay on the line is computed numerically by means of the summation of all the individual delays on the line. All the individual delays are computed by equation (3), subject to non-negativity.

Table 3 - Individual delays [min] on a railway line, given a primary delay [min]. Positive delays are bolded; color intensity proportional to cell values. Stations are also given a letter ID.

		STATION							
		1	2	3	4	5	6	7	8
		A	B	C	D	E	F	G	H
TRAIN	1	**44**	**42**	**40**	**38**	**36**	**34**	**32**	**30**
	2	**38**	**36**	**34**	**32**	**30**	**28**	**26**	**24**
	3	**32**	**30**	**28**	**26**	**24**	**22**	**20**	**18**
	4	**26**	**24**	**22**	**20**	**18**	**16**	**14**	**12**
	5	**20**	**18**	**16**	**14**	**12**	**10**	**8**	**6**
	6	**14**	**12**	**10**	**8**	**6**	**4**	**2**	0

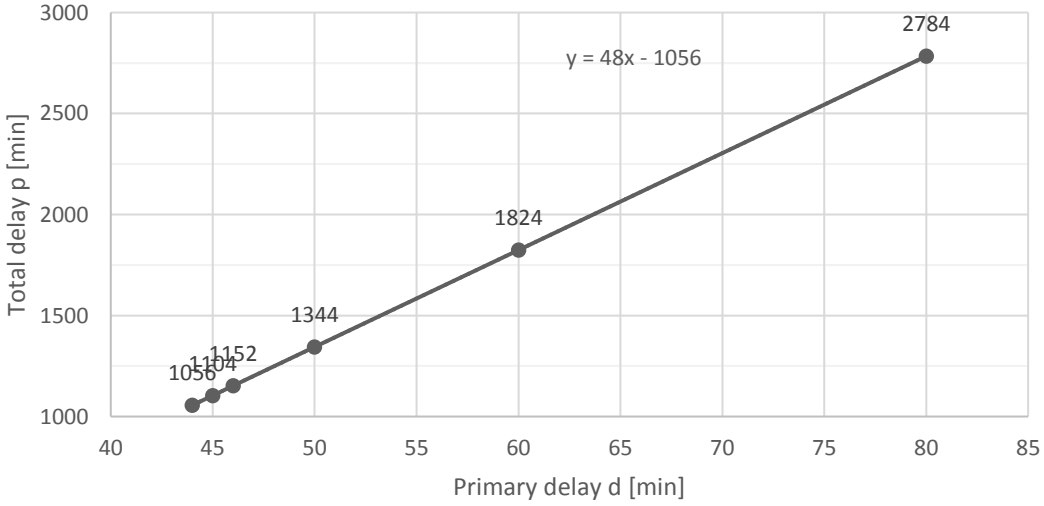

Figure 1 - Total delay measured on the line as a function of the primary delay given to the first train at the first station

The total delay measured at different values of primary delay is reported in the chart above: it represents a linear relation, with the specific formulation.

A case with recovery

The same example as before can represent the case with complete recovery for primary delays up to *p=14* minutes. This case is represented in the table below.

Table 4 - Individual delays [min] on a railway line, given a primary delay [min]. Positive delays are bolded; color intensity proportional to cell values. Stations are also given a letter ID.

		STATION							
		1	2	3	4	5	6	7	8
		A	B	C	D	E	F	G	H
TRAIN	1	**14**	**12**	**10**	**8**	**6**	**4**	**2**	0
	2	**8**	**6**	**4**	**2**	0	0	0	0
	3	**2**	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0

The following chart represents the total delay computed both as the numerical summation of the individual delays and through the analytical formulation (10).

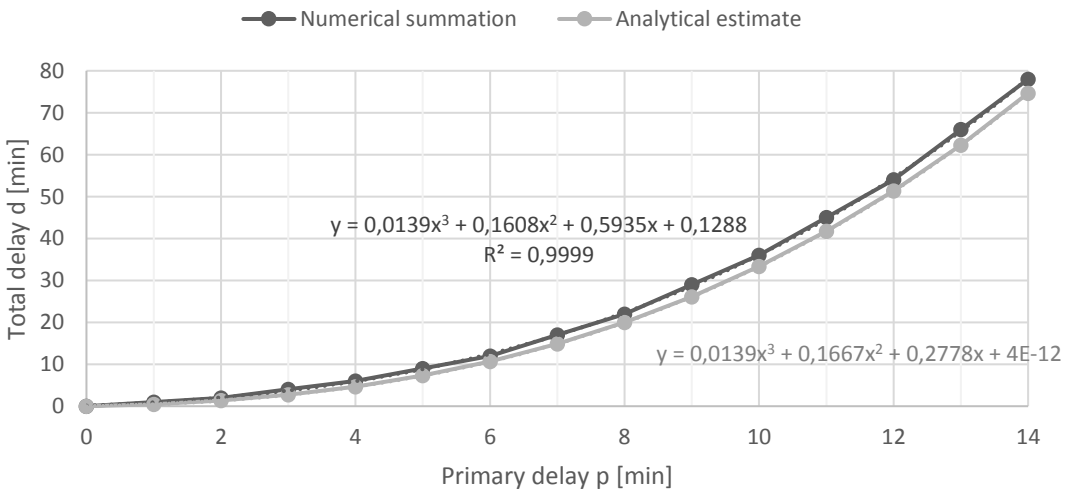

Figure 2 - Total delay measured as the numerical summation and estimated through the analytical formulation.

The numerically computed total delay is regressed to a cubic polynomial expression, given in the chart, along with the R^2 index, which shows a high quality regression. The equation of the analytical total delay is given as well.

The general case

The results of previous examples are part of a composite function, which borders are be defined here below. The cubic section extends in the range of primary delay $p \in [0,16]$. The linear range, on the opposite side, starts at the value $p > 44$ *minutes*.

It is worth to notice that the intermediate range can be regressed perfectly to a quadratic parabola, consistently with the assumed shape in the theoretical explanation. The regression is shown in the picture below.

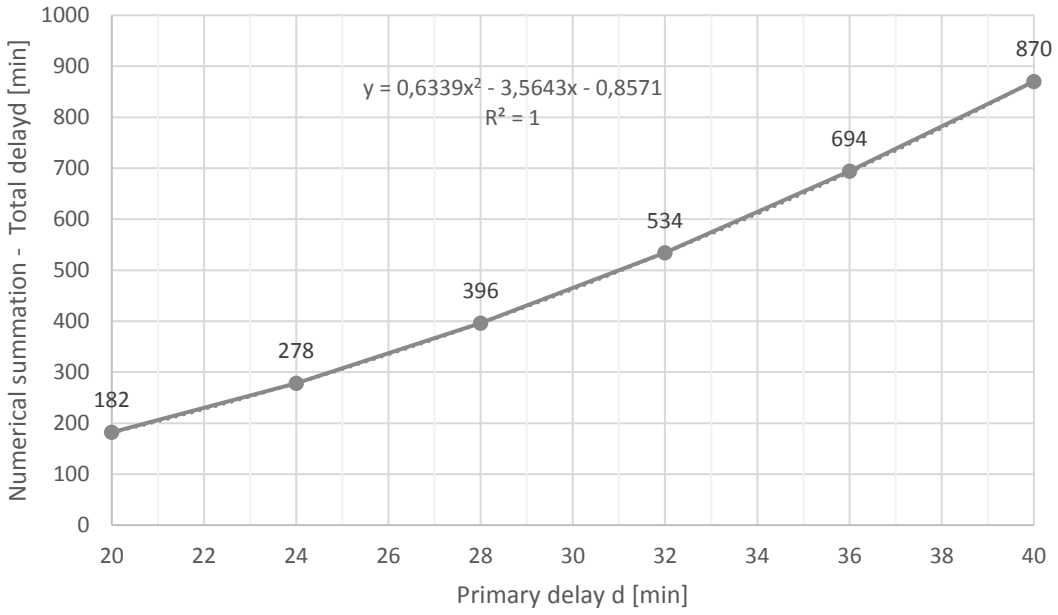

Figure 3 - The total delay numerically summed in the example. Partial recovery range.

The chart below shows the three segments of the total delay as a function of the primary delay with the equations and coefficients of the regressed polynomial functions.

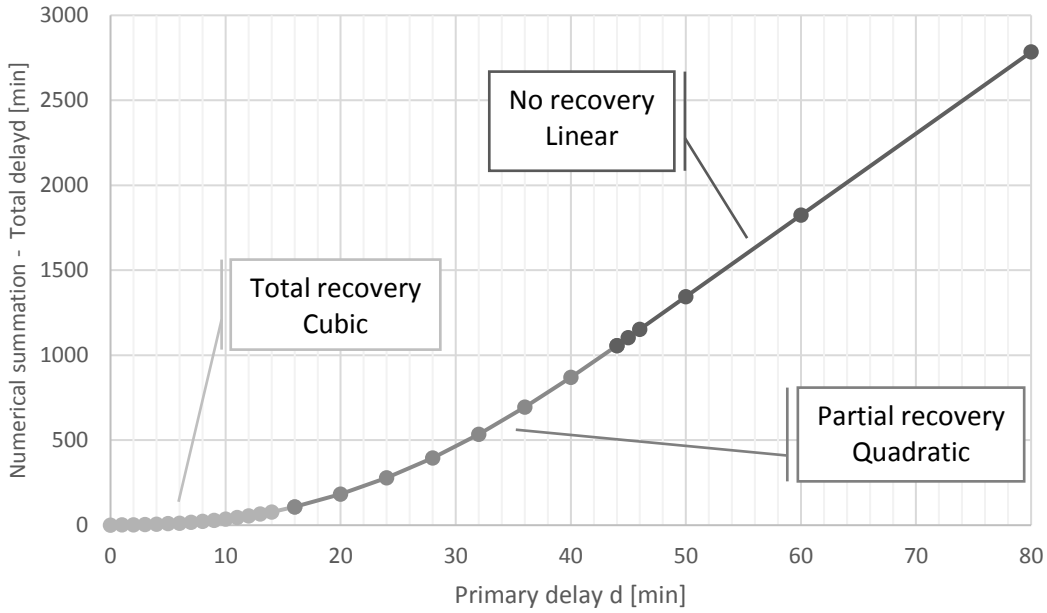

Figure 4 - Total delay summed in the numerical example.

5.2. A real large scale case: the Nordbane in Copenhagen

The numerical examples above were subject to a number of assumptions that could be too strong in some cases. The assumption of uniform timetable allowance and buffer time, in particular, could be unrealistic for some railway lines. This is compared with a real case in the following, where measurements of the total delay on a simulated railway line, with no information about the supplement times and the buffer between trains.

To validate the polynomial shape of the total delay against the primary delay we simulated the operation of a suburban railway line in Denmark. The suburban railway network in Copenhagen is a very densely occupied network with 2 minutes headway in the busiest section. Six different lines operate on the network, five out which run through the same central section. The suburban line is operated by uniform rolling stock in cyclic timetable. The selected section of the suburban network is the line from Hellerup to Hillerød. Overtakes in this section are prevented. Though it is theoretically possible at selected stations, it hardly occurs in real operation, due to the very high frequency of the train service.

The micro-simulation software RailSys by Rail Management Consultants GmbH (RMCon) was used for the simulation. This micro-simulation uses continuous computation of train motion equations and simulates the interaction between trains through discrete processing of signal box states. Given user defined infrastructure, rolling stock, and timetable databases, it is possible to calibrate the train paths defining the running time supplements; moreover, different driving behaviors can be modelled for on time trains and delayed ones. The strength of the micro-simulation models is the higher accuracy than the analytical models, and their flexibility to represent different contexts. Changes in the infrastructures and operating rules can easily be implemented and tested. The accuracy comes, though, at the cost of much longer computation time, as well as set-up time. Other micro-simulation software is available on the market, like OpenTrack by OpenTrack Railway Technology Ltd. Despite some differences in the approach, both the mentioned software suffer from long time needed to compute such detailed models (Landex, 2008).

The model formulated in the paper can, on the other hand, be interpreted as a macroscopic simulation tool, which only has basic information on the link between two stations and the constraints in operation. The only pieces of information required in this model are the buffer times and the timetable allowance. We used a microsimulation tool to validate this in a more detailed model, which includes data about the geometry of the infrastructure and the actual separation of trains through the signaling system.

Table 5 - Stopping patterns of each train path. X = Stop, | = pass

Station	**Code**	**km**	**A**	**E**
Hellerup	Hl	7,8	X	X
Bernstorffsvej	Btf	9,3	|	X
Gentofte	Gj	10,9	|	X
Jægersborg	Jæt	12,6	|	X
Lyngby	Ly	13,9	X	X
Sorgenfri	Stf	15,9	|	X
Virum	VG	17,7	|	X
Holte	Hot	19,0	X	X
Birkerød	BG	23,8	X	
Allerød	LG	29,3	X	
Hillerød	HG	36,5	X	

Two different train paths run every ten minutes on the line between Hellerup and Lyngby with two different stopping patterns:

- Line A: runs throughout the entire line, skipping 5 stops in the first stretch
- Line E: only runs the first stretch, stopping at all the stations.

The line stationing and stopping patterns are summarized in Table 5.

We Monte Carlo sampled *n=100* values of primary delay from a uniform distribution between 0 and 10 minutes and measured the related total delay developed on the line. Delays up to ten minutes were completely recovered by all the trains before the end of the line, meaning that this case is with complete recovery. The resulting shape of the total delay against the primary delay is a cubic parabola.

The same primary delay given to trains from different lines resulted in different values of total delay. The graph below collects the total delay values for primary delays given to either line A or E. The graph shows that there is not a unique curve that represents the total delay for any train delayed in the timetable. The total delay keeps cubic against primary delay given to individual lines, and two different regressions are applied to the results.

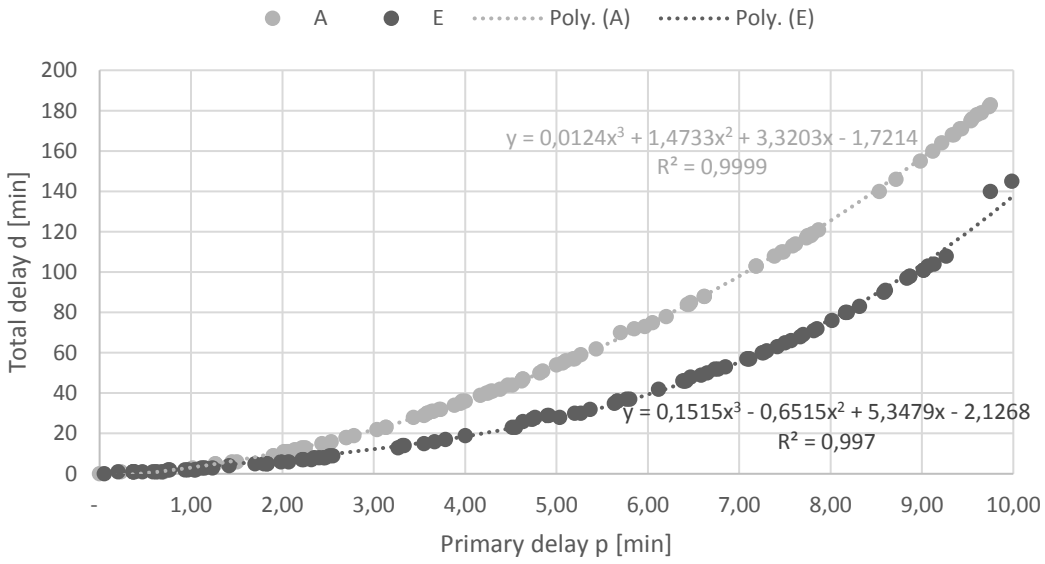

Figure 5 - Measured total delay on the suburban railway line Hellerup - Hillerød, as a function of the primary delay. Each curve represents the total delay generated delaying either railway line of the timetable.

5.3 Discussion

The numerical examples revealed a high accuracy of the model for both the linear segment and the cubic one. The estimation of the total delay in the case with no recovery is exact, and shows the linear relation with the primary delay. The estimation of the total delay presents, on the other hand, an approximation in the range with total recovery. This is due to the upper bound of the summation, which is integer by nature but is approximated continuous in the summation solution.

The microsimulation of a suburban railway line revealed that the proposed model is not limited to simplified theoretical lines. The simulated railway line is characterized by non-uniform timetable allowance and stations not equally spread along the line. Moreover, we simulated heterogeneous stopping pattern and non-uniform buffer time between trains on the line. The trains also had different scheduled running times in the segments Hellerup-Lyngby and Lyngby-Holte. Nevertheless, the total delay shape against the primary delay is still close to a cubic parabola. Delays up to 10 minutes were completely recovered by all the trains before the end of the line, in accordance with the model developed. This shape validates the polynomial expression of the total delay as a function of the primary delay for small primary delays.

It is worth to notice that the graph also shows some phenomenon that is not fully explained by this paper. The total delay generated by the same amount of primary delay is different, depending on the first train delayed. This is due to the different interactions that are generated. The primary delay is generated at the first station. When the delay is given to a train from line A, it propagates to both the following stations and the following train. The A line is faster than the E line, so the buffer time between a train A and a train E is minimum at the first station and increases on the way. The opposite is valid in the reversed sequence: the buffer time between a train E and a train A is maximum at the first station and decreases to the minimum at Holte, which is the last station of line E. The delays given to a train E do not propagate to the following train until a value that is large enough, so that the train E hinders the following train A at Holte station. This value equals the total timetable allowance of train E from Hellerup to Holte, increased by the buffer time

between line E and line A at Holte. This phenomenon is the reason why the total delay values related to primary delays given to line E are smaller than line A.

6. CONCLUSIONS AND CONTRIBUTIONS

In this paper, we presented a model of the total delay on railway lines as a composite polynomial function of the primary delay given to one train. The function can be divided into three sections, which boundaries are specific values of the primary delay. The values can be computed under the assumption of uniform timetable allowance and buffer time, and given the number of stations on the line and the traffic volume of the timetable.

Despite the different approach, our results are consistent with Hasegawa et al.'s findings: their fluid model returns the total delay as a cubic polynomial function of the primary delay assuming complete recovery. We found a cubic relation still valid, when considering the trains as discrete separate individuals.

Our estimation method presents an approximation, due to the discrete nature of the upper boundaries in the summations. This approximation leads to a small underestimation of the total delay in the cubic section, and further research development is needed to estimate an expression for the error. The approximation expires in the linear section of the total delay, which is computed exactly. The same kind of error is found and tackled by Landex, though he tackles the problem only looking at the buffer times between trains. He solves the issue considering only the absolute value in the ratios that define the summation boundaries. In this way, the simplicity of the formulation is sacrificed to improve the estimation accuracy. Further investigation could clarify if the error reduction justifies such a complication of the model.

We validated the theoretical model through a numerical example with ideal conditions: equal train paths, uniform timetable allowance and buffer times over trains and station. The case study demonstrated that even in real conditions that differ from the hypothesis of the model, the total delay is still a cubic function of the primary delay.

This model lays the basis for future development of stability analysis of suburban railway line with limited use of micro-simulation, as well as metro-like systems, which typically have homogenous timetables and rolling stock. This overcomes the microsimulation's problems, listed by Mattsson and Landex, reducing the time needed for this kind of analyses, and opening the way for real time applications. The estimation of the total delay, given the value of primary delay gives a measure of the effect related to a disruption, which is namely the stability. The model can be extended to railway networks, linking different suburban railway lines together. The linkage should be the objective of studies to set the validity limits of the extension from lines to networks.

The advantage of such a model in railway networks and hence the contribution of this paper is in its simplicity and analytic formulation and closed form expression, compared to micro-simulation models, yet giving and accurate estimation of the total delay. Stability analyses typically deal with small amount of primary delays, so only the cubic section of the model is needed, as shown in the case study Hellerup – Hillerød. Nevertheless, the intermediate segment could still be formulated from a theoretical point of view: a quadratic polynomial regression resulted in the numerical example. This will make it possible to make much faster initial analyses of punctuality of timetable in the initial strategic planning phases when designing new timetables.

The average timetable allowance and buffer time can also be computed by reversing the total delay formulation, due to the close-form expression of the model. This means that with a given desired punctuality of the railway line, the needed timetable supplement can be estimated by the function. This would also ease the planning process compared to a try-and-correct approach using microsimulation.

7. BIBLIOGRAPHY

Carey, M. (1999). Ex ante heuristic measures of schedule reliability. *Transportation Research Part B: Methodological*, *33*(7), 473–494. http://doi.org/10.1016/S0191-2615(99)00002-8

Cerreto, F. (2015). Micro-simulation based analysis of railway lines robustness. In *6th International Conference on Railway Operations Modelling and Analysis* (pp. 164–1 –164–13). Tokyo: International Association of Railway Operations Research. Retrieved from http://orbit.dtu.dk/fedora/objects/orbit:140777/datastreams/file_112408001/content

Goverde, R. M. P. (2010). A delay propagation algorithm for large-scale railway traffic networks. *Transportation Research Part C: Emerging Technologies*, *18*(3), 269–287. http://doi.org/10.1016/j.trc.2010.01.002

Goverde, R. M. P., & Hansen, I. A. (2013). Performance indicators for railway timetables. In *2013 IEEE International Conference on Intelligent Rail Transportation Proceedings* (pp. 301–306). IEEE. http://doi.org/10.1109/ICIRT.2013.6696312

Hasegawa, Y., Konya, H., & Shinohara, S. (1981). MACRO-MODEL ON PROPAGATION-DISAPPEARANCE PROCESS OF TRAIN DELAYS. *Railway Technical Research Institute, Quarterly Reports*, *22*(2), 78–82. Retrieved from http://trid.trb.org/view.aspx?id=180725

Kirchhoff, F. (2014). Modelling Delay Propagation in Railway Networks (pp. 237–242). Springer International Publishing. http://doi.org/10.1007/978-3-319-07001-8_32

Landex, A. (2008). *Methods to estimate railway capacity and passenger delays*. Technical University of Denmark (DTU). Retrieved from http://findit.dtu.dk/en/catalog/2185768953

Mattsson, L.-G. (2007). Railway Capacity and Train Delay Relationships. *Critical Infrastructure: Advances in Spaital Science*. Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-68056-7_7

Meester, L. E., & Muns, S. (2007). Stochastic delay propagation in railway networks and phase-type

distributions. *Transportation Research Part B: Methodological*, *41*(2), 218–230. http://doi.org/10.1016/j.trb.2006.02.007

Pellegrini, P., Marlière, G., & Rodriguez, J. (2014). Optimal train routing and scheduling for managing traffic perturbations in complex junctions. *Transportation Research Part B: Methodological*, *59*, 58–80. http://doi.org/10.1016/j.trb.2013.10.013

Radtke, A. (2008). Infrastructure Modelling. In I. A. Hansen & J. Pachl (Eds.), *Railway timetable & traffic* (1st ed., pp. 43–57). Hamburg: Eurailpress. Retrieved from http://www.iaror.org/documents/TimetableandTraffic2008ab.pdf

Salido, M. A., Barber, F., & Ingolotti, L. (2008). Robustness in railway transportation scheduling. In *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)* (pp. 2833–2837). Chongqing, China: IEEE. http://doi.org/10.1109/WCICA.2008.4594481

Schittenhelm, B. (2011). Planning With Timetable Supplements in Railway Timetables. In *Annual Transport Conference at Aalborg University*. Aalborg, DK: trafikdage. Retrieved from http://www.trafikdage.dk/papers_2011/63_BerndSchittenhelm.pdf

Siefer, T. (2008). Simulation. In I. A. Hansen & J. Pachl (Eds.), *Railway timetable & traffic* (1st ed., pp. 155–169). Hamburg: Eurailpress. Retrieved from http://www.iaror.org/documents/TimetableandTraffic2008ab.pdf

Siefer, T., & Fangrat, S. (2012). Effects of Shifting Running Time Supplements. In *Proceedings of the 1st IWHIR* (Vol. 1, pp. 433–440). Berlin.

Takeuchi, Y., & Tomii, N. (2005). Robustness Indices for Train Rescheduling. In *1st International Conference on Railway Operations Modelling and Analysis RailDelft2005* (pp. 1–19). Delft, The Netherlands.

UIC. (2000). *UIC leaflet 451-1 Timetable recovery margins to guarantee timekeeping - Recovery margins* (4th ed.). Retrieved from http://www.uic.org/etf/codex/codex-detail.php?langue_fiche=E&codeFiche=451-1

Vromans, M. J. C. M., Dekker, R., & Kroon, L. G. (2006). Reliability and heterogeneity of railway services. *European Journal of Operational Research*, *172*(2), 647–665. http://doi.org/10.1016/j.ejor.2004.10.010

Wiklund, M. (2002). *The vulnerability of the railway transport system - A structure for formulation of models and development of methods*. *VTI meddelande* (Vol. 932). Retrieved from http://www.vti.se/sv/publikationer/jarnvagstransportsystemets-sarbarhet---struktur-for-modellformulering-och-metodutveckling/