Quantifying the effects of running time variability on the capacity of rail corridors*

Norman Weik[1], Nils Nießen

Institute of Transport Science,
RWTH Aachen University,
Mies-van-der-Rohe-Str. 1,
52074 Aachen, Germany

[1] E-mail: norman.weik@rwth-aachen.de, Phone: +49 (241) 80 25185

Abstract
Traffic variability is well known to have a substantial effect on railway capacity. Varying train running and stopping times entail larger train separations and yield non-usable time slots in train timetables and operations. In this paper, we aim to assess the effects of uncertainty in running and dwell times on the capacity of railway corridors in long term planning of rail traffic. Our main focus are commuter and metro systems, where the effects of fluctuations of running and stopping times are particularly pronounced due to dense operations. To analyze the effects of variability in train operations, we propose a new stochastic approach based on a serial queuing network with finite capacity service stations. The corridor is modeled as a sequence of heavily correlated service stations representing line segments and stations, for which effective throughput, distributions of train running times and service quality are calculated. The performance of the model is tested in a case study for the central link of the mass transit system in Cologne. In addition, an outlook on how the model can be extended to general heavy-rail corridors with different types of train services is provided.

Keywords
capacity, railway corridor, long term planning, variability, queuing network

1 Introduction

Planning for the unplanned is arguably one of the most challenging aspects in railway operations research. Uncertainty and randomness enter in the form of varying running times and delays in train operations or differing train path requests, train sequences and headways in the capacity allocation process.

Long term planning, where strategic decisions with long-lasting effects on the appearance of rail networks are taken, is particularly susceptible to variations in traffic evolution forecasts. Due to the infrastructure's long renewal cycles of the order of 30 years, traffic demand, driving characteristics and timetable structure will almost surely change within its

*This is the Authors' Accepted Manuscript of the following article: N. Weik, N. Nießen, Quantifying the effects of running time variability on the capacity of rail corridors, Journal of Rail Transport Planning & Management 15, 100203, 2020, which has been published in final form at https://doi.org/10.1016/j.jrtpm.2020.100203

designated lifetime. At the same time, infrastructure design is accompanied by high investment costs and hence poses a substantial financial risk to the infrastructure manager. For this reason, the infrastructure should be robust against deviations from the envisioned traffic concept and flexible and performant enough to allow for changes in timetable structure. It is this strategic long-term planning view on infrastructure and line concepts, which is independent of the exact train sequence in the timetable, we adopt in this paper.

In this context, we investigate the effects of traffic variability on rail capacity. It is well known that variations of train running times, station stopping times, driving characteristics, and headways tend to have a negative effect on rail capacity (UIC, 2004; Dingler et al., 2013). This is all the more true if the traffic concept itself is subject to uncertainty and hence contributes to the heterogeneity of rail traffic in the sense of varying line frequencies and train headways.

Presently, most methods in capacity analysis have been focusing on either service or timetable variability: The effects of stochastic train running times have been studied using algebraic approaches in (Goverde, 2007; Huisman and Boucherie, 2001; Meester and Muns, 2007). Similarly, delay propagation modeling with stochastic running times or in view of random perturbations has been reported (Yuan and Hansen, 2007; Büker and Seybold, 2012). These models generally build on a fix train order as the number of possible train combinations in case of random timetables becomes untractable.

Another aspect of service variability is traffic heterogeneity. In the field of capacity analysis, traffic heterogeneity is generally reviewed in its effects on infrastructure occupation (UIC, 2004) or a-posteriori simulations (Warg, 2012; Dingler et al., 2013), where the sum of reciprocal headways (SSHR) has been introduced as a measure of heterogeneity (Vromans, Dekker, and Kroon, 2006). Timetable variability is accounted for by ensemble averaging techniques (Jensen et al., 2017; Lindfeldt, 2010) or with probabilistic approaches to train order and headways and a deterministic or statistic description of train running times (Schwanhäußer, 1994). Queuing models, which are naturally suited to cope with both random timetables and service processes, suffer from the fact that they can only be applied locally (Schwanhäußer, 1994; Wendler, 2007) or make simplified assumptions on service time statistics and station capacity (Huisman, Boucherie, and van Dijk, 2002). What is more, they generally rely on an abstracted representation of the infrastructure rendering the localization of infrastructure effects extremely difficult.

In the present paper we show how variability effects can be investigated on the system level in a timetable independent a-priori model. To this end we introduce a new queuing network approach to model rail corridors, which draws from techniques used to model production lines in manufacturing system analysis. Directional track segments are viewed as finite buffer work stations in a service network. By modeling block sections or track segments as finite capacity queuing stations, network correlations such as the spillback of delays resulting from capacity limitations can be accounted for. This allows to pinpoint infrastructure bottlenecks as well as their effects on the traffic situation on the corridor. In addition, the effectiveness of infrastructure adjustments with respect to the overall network performance can be studied.

While still being hard to solve for general service processes, efficient numerical approaches for performance analysis of this type of systems have been discussed in the context of manufacturing systems (cf. Curry and Feldman (2011) for an introduction to the topic). It has been shown that the negative effects of variations on capacity can be understood by two main factors: Blocking and starvation. Blocking refers to the fact that a request finishing

its service at a service location cannot proceed to the next location as the next server is currently dealing with another request and no storage capacity is available. Starvation denotes the situation that a service station is forced to idle as requests can not be supplied on time by the upstream queue, such that precious service capacity is wasted.

Building on previous work by Buzacott, Liu and Shanthikumar (1995), Tempelmeier and Bürger (2001) and, more recently, Bierbooms, Adan and van Vuuren (2013), both the effective running times and maximum achievable throughput including the propagation of network effects due to service time variation can be calculated. It is important to point out that the achievable throughput does not correspond to the absolute or theoretical capacity known from deterministic modeling of railway capacity, but yields a smaller value which already takes into account the effects of variations.

In this paper we demonstrate the use of the model in traffic planning for rail corridors. It is shown how the methodology can be used to determine the maximum number of trains that can be operated on a corridor such that a) the system remains stable in view of varying train running and station dwell times and b) the effective running time remains bounded by a certain threshold value given correlations are accounted for. The focus of this paper is on heavily loaded central corridors of metro transit systems. Here, the question of the maximum feasible frequency of trains that can be operated on the corridor is central – especially for traffic planning during peak hours. We test our methodology in a case study based on the central commuter rail corridor in Cologne, Germany.

Nonetheless, the method is not limited to metro transit applications. In an extended outlook we also explore an option to use the model in the context of general heavy rail corridors with different types of train services. Here, significantly differing train speeds and the possibility to swap train order in stations introduce additional complexity.

Apart from providing a more detailed understanding of the effects of variability on the available capacity, our methodology allows to perform a structural investigation of the railway infrastructure. It provides the means to detect bottlenecks resulting from collaborative effects between different network elements, which – individually – may be subcritical. The effective train running times established with the method and their deviation from input running times can also be used to derive an approximation of delay build-up or (in case minimum technical running times are used as input) as an indication on the optimal allocation of buffer times and timetable supplements.

In the following section we start by briefly reviewing the literature on railway capacity analysis, focusing on previous work on variability as well as queuing models for rail traffic modeling. Our model as well as its solution is described in Section 3. The functionality of our method is demonstrated in a case study based on the central link of the Cologne commuter rail system. Possibilities to extend the model to more general rail corridors with different train types and overtakings are discussed in Section 5.

2 Literature Review

2.1 Railway Performance Modeling

The main goal of railway capacity planning is to assess the number of trains that can be operated on a given part of the railway infrastructure within a certain time period. Different capacity metrics, such as absolute capacity, denoting the maximum number of trains that can theoretically be scheduled, the occupation ratio (i.e. the time share the infrastructure

is occupied) or the height and probability of delays have been used (Abril et al., 2007). While the absolute capacity is relatively easy to access analytically, it is is usually hardly achievable in railway systems, which is why the service level dependent practical capacity based on quality metrics such as delays and occupation is much more widespread.

In medium term tactical planning, capacity analysis most often relies on an existing operational concept or even a fully constructed timetable. Optimization approaches for assessing the absolute capacity of railway networks have been proposed by Burdett and Kozan (2006) for general, non-periodic timetables. For periodic timetable a max-plus approach to calculate mean cycle lengths has been discussed by Goverde (1998). The same approaches can also be used to determine whether a feasible timetable can be generated and the amount of slack that can be incorporated in a timetable. Cacchiani, Caprara and Toth (2010) provide a time dependent routing problem to schedule extra freight trains in an existing passenger timetable concept.

Timetabling approaches have also been discussed in conjunction with service dependency. Goverde (2007) proposes a max-plus technique to assess the mean cycle length of a periodic timetable in the presence of delays progressing in the network. Burggraeve and Vansteenwegen (2017) aim for an analysis of timetable robustness and an optimized allocation of timetable reserves.

Arguably, the most generally used method of capacity analysis is the timetable compression method according to UIC 406 standard, where the occupation ratio of timetables is assessed (UIC, 2013). Various adaptations have been discussed, e.g. for application to stations and junctions or to incorporate overtakings within the line segment to avoid underestimating capacity utilization due to the negligence of correlations (Landex et al., 2006; Kuckelberg, Gröger, and Wendler, 2011). While the method is defined based on infrastructure utilization as a capacity metric it has also been used synthetically in timetable saturation approaches to determine the residual number of trains that can be scheduled (Lindner, 2011).

For long term planning and infrastructure planning, in particular, timetable based approaches are not optimally suited, as either no detailed timetable concepts are available or – in case they do – the latter are subject to sincere modification during the typical lifetime of railway infrastructure. This is why the picture provided by a purely timetable based approach – while being more exact for a single timetable – remains incomplete. By investigating a single timetable only, infrastructure bottlenecks tend to be masked by a timetable concept which is optimally designed to accommodate for theses shortcomings.

Ensemble averaging has been introduced as a workaround for this problem. Jensen et al. (2017) have recently provided a graph theoretical modeling of train interactions based on sectional minimum headway times and provided an algebraic approach to calculate the mean cycle length, and hence the infrastructure occupation both locally and network-wide. Unlike other algebraic techniques, such as the approaches in Büker and Seybold (2012) or Goverde (2007), the approach provided by the authors does not require a given train order, but performs a distributional analysis of timetable realizations based on Monte-Carlo sampling. A similar idea has been pursued by Lindfeldt (2010) in the so-called Timetable Variant Evaluation Approach (TVEM), where a sample of timetable variants has been investigated using a posteriori traffic simulations with RailSys to calculate average delays for a given infrastructure.

Stochastic models have also been used in this context. Yuan and Hansen (2007) and Büker and Seybold (2012) discuss delay propagation models to calculate arrival punctuality and knock-on delays based on distributional assumptions on buffer times and initial delays.

While no specific timetable is required, train order is assumed to be given. The variability of driving parameters such as running times and minimum headway times, however, is not accounted for. A generalization to general running times is however possible and has been described by Huisman and Boucherie (2001) and Meester and Muns (2007).

Another class of capacity models reflecting variability effects are queuing methods. Queuing based approaches to assess the required number of station tracks based on train waiting probabilities have been discussed by Potthoff (1962); Hertel (1984) and Landex (2011). For line sections and route nodes queuing based approaches to assess the scheduled waiting times have been developed by Schwanhäußer (1994) and Wendler (2007). Knock-on delays for both line segments (Schwanhäußer, 1974; Weik, Niebel, and Nießen, 2016) and station threads (Nießen, 2013) have been assessed in a delay propagation model adopting a queuing system perspective of the railway infrastructure and using queuing theory results to extrapolate heavy traffic limits. Weik and Nießen (2017) recently introduced a generalization, where service times are modulated according to the current state of the infrastructure.

Most queuing models presently used in the railway context have been used in a local setting for the analysis of individual line segments, station threads or route nodes. Huisman, Boucherie, and van Dijk (2002) discuss a queuing network approach to analyze an entire railway subnetwork. While being restricted to exponential service times for solvability, the authors propose an expansion technique to model lines as a series of M/M/1-queues, such that the free (unimpeded) phase-type distributed running times on line segments match the first two moments of given running time statistics.

Queuing network models with more general service processes can rarely be solved analytically due to correlations between service stations. The same holds true for finite capacity queuing networks, where the waiting area in front of service stations is limited. Yet, the restriction to infinite capacity server stations limits the usability of queuing network approaches as spillbacks of delays in the network cannot be modeled. Osorio and Bierlaire (2009) have provided an approximation scheme for road traffic, where the stationary distribution and the blocking parameters of the queues are solved consistently as a nonlinear system of equation. However, the model is unable to cope with non-exponential service times.

Additional insights can be gained from queuing network models used in the context of manufacturing systems. Here, production lines with general servers and finite capacity have successfully been solved using numerical approaches. Two main approaches can be distinguished: Decomposition methods (Gershwin, 1987), where the production line is decomposed into a series of tandem server queues, which – individually – can be solved analytically. Correlations are accounted for in an iterative forward-backward updating procedure until the queue performance parameters in the production line are consistent. The second approach is an expansion technique introduced by Kerbache and Smith (1987). Here, finite capacity stations are expanded by a dedicated holding node to which blocked requests are routed probabilistically.

An important conclusion that can be drawn from queuing network models used in this area is the fact that, on the system level, the maximum capacity in the presence of variability is far lower than in deterministic systems, where the slowest queue determines the achievable throughput. To the best of our knowledge, no a-priori investigation of the effects of service time variability on the maximum achievable capacity in railway systems on the system level has been undertaken.

3 Method

3.1 Queuing Network Models

In the present paper we explore queuing based modeling of railway systems. Similar to the approach by Huisman, Boucherie, and van Dijk (2002), railway corridors are represented by a series of queues, where service times are fitted to train running times on the corresponding infrastructure. However, we do not pursue a Jackson network approach, where each queue is modeled as an $M/M/1$-queue with infinite capacity. In our model, queues are allowed to have general arrival and service processes and finite capacity, such that the physical infrastructure restrictions on train access are reflected in the topology of the model. A correspondence of block sections and queuing stations is incorporated by taking block sections to be $G/G/1/0$-queues, such that blocking and spillback are accounted for. The method we propose is capable to treat feed-forward type networks, i.e. networks without cycles, such that directionality can be defined in the network. This allows to investigate rail corridors and many urban transit networks.

Correlations between queues naturally arise in case of non-exponential holding times, which makes general queuing networks hard to solve and often only numerically or approximately tractable. For general queuing networks the arrival process of a queue is correlated with the service process of its predecessor queue. This requires an explicit modeling of

- starvation, where Queue i is depleted and forced to idle until the next request has finished service at predecessor queue $i-1$.
- Additionally, finite capacity entails blocking. That means that a request finishing its service at Queue i cannot transfer to Queue $i+1$, because the buffer in front of the successor station is full or – in case the successor station has buffer size 0 – the successor queue is currently still serving another train.

Blocking and starvation render the effective throughput of queues within a network context smaller than the corresponding throughputs of the queues in isolation. This fundamental queuing theory result is well known in the context of railway capacity under the term "network effects". In Landex et al. (2006), for instance, the influence of the observation area has been investigated for the Danish railway network. Delays of a railway line were found to be almost 72% higher in case the entire Danish railway network was considered as compared to the case where the line was analyzed in isolation.

Due to their stochastic nature queuing-based models are unsuited to provide insights into situations where highly structured (periodic) timetables with detailed connectivity information are prevalent. Still, they allow to investigate the system's general behavior as well as to identify infrastructure bottlenecks, which makes them particularly suited for long-term planning purposes.

3.2 Modeling of Railway Corridors

For the modeling of double-track railway corridors we assume service stations in the queuing network correspond to block sections on the infrastructure, i.e. we consider a linear network of $G/G/1/0$-queues. The two directional tracks of the corridor are modeled as two independent queuing lines as track changes are not performed in regular operations due to the negative implications on capacity.

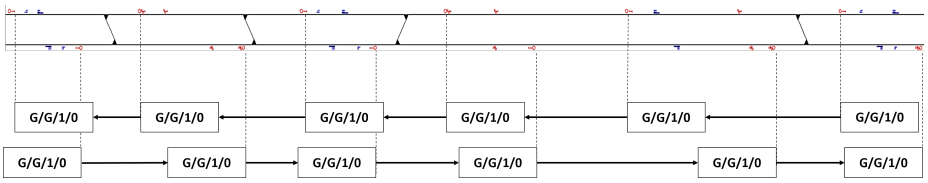

Figure 1: Modeling of railway infrastructure as a linear queuing network.

Metro transit systems, that we are primarily interested in in this paper, typically exhibit similar train patterns, stopping policies and driving characteristics. As a result, we can restrict ourselves to a single customer class.

Service times of the single queues correspond to train running times in the corresponding block. Strictly speaking, simultaneous occupations of neighboring blocks to model approach and clearance times would have to be considered (Pachl, 2014). These effects can theoretically be incorporated in the model by resorting to service stations with correlated arrival and service process – which can be either aborded analytically (cf. Mitchell and van de Liefvoort (2003)) or via complicated phase-type distributed service and arrival processes as in Bierbooms, Adan and van Vuuren (2013). As the main source of variation as well as the critical elements in mass transit systems are station stopping times, the effects of this approximation generally are not that pronounced.

3.3 Solution Technique

Overview

$G/G/1/N$-feed forward queuing networks can be solved with Gershwin (1987)'s decomposition technique. The basic idea consists in decomposing the network into a sequence of $N-1$ tandem queuing systems, i.e. queuing networks consisting of two servers and one buffer between the two servers (which can have size 0 for queues without waiting capability). This simple queuing network is well understood and can be solved exactly (cf. Gershwin (1987)). The departure server of subsystem i equals the arrival server in subsystem $i+1$ (see Figure 2). The departure server of subsystem i is blocked if the buffer (or the next queue for zero-storage queues) in system $i+1$ is full at the instant of a service completion. Equally, the arrival server of system $i+1$ is starved if the buffer in system i has run empty once the last customer completes its service in subsystem $i+1$.

To correctly catch the dependencies between the servers in the network and to analyze blocking and starvation effects, a backward-forward iteration scheme has been introduced by Gershwin (1987). It consists of two steps:

- A forward propagation pass of the network, where the service rate of the arrival server in subsystem $i+1$ is updated by the idling rate of the departure server in i. This accounts for starvation as the starvation probability of arrival server A_{i+1} is given by the probability that server D_i idles.
- A backward propagation pass, where the service rate of the departure server in subsystem $i-1$ is prolonged by the blocking probability and duration of subsystem i. The blocking probability of system i denotes the probability that buffer $i \rightarrow i+1$ is full upon service completion at server D_{i-1}.

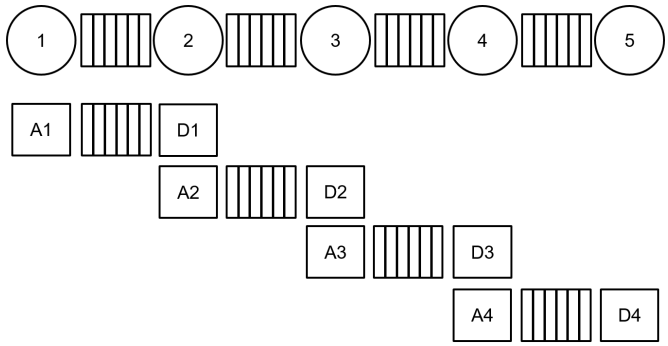

Figure 2: Decomposition approach for finite capacity serial queuing networks. The series of N queues is decomposed into $N-1$ tandem server subsystems.

After each updating of the input parameters of a queue its queue length distribution and the performance parameters, i.e. blocking and starvation probability p_b and p_s, as well as the effective throughput TH of the subsystem, are updated. The forward-backward passing procedure is repeated until convergence of the line's throughput is achieved. An schematic overview of this well-known forward-backward algorithm is given in Algorithm 1.

Initialization;
Initialize departure and arrival server. ;
Set mean service rates and coeff. of variation to values for corresponding queue (i.e. $p_s = 0$, $p_b = 0$, and $TH = 1/D(1)$ are assumed at the start);
Iteration;
while $|TH - TH_{old}| > \epsilon$ **do**
  $TH_{old} = TH$;
  for $i = 2 : n_{queues}$ **do**
    $A(i) = S(i) + \underbrace{\frac{1}{TH(i-1)} - D(i-1)}_{starvation}$;
    update $c_A^2(i)$;
    $\rho(i) = D(i)/A(i)$;
    Calculate new $TH(i)$ based on queue length distribution of subsystem i;
  end
  for $i = n_{queues} - 1 : 1$ **do**
    $D(i) = S(i) + \underbrace{\frac{1}{TH(i+1)} - A(i+1)}_{blocking}$;
    update $c_D^2(i)$;
    $\rho(i) = D(i)/A(i)$;
    Calculate new $TH(i)$ based on queue length distribution of subsystem i;
  end
end

Algorithm 1: Backward-Forward Propagation Algorithm.

$S(i)$ denotes mean service time of Queue i, $D(i)$, $A(i)$, $c_A^2(i)$ and $c_D^2(i)$ are the mean and the squared coefficient of variation (SCV) of the service time of the departure and arrival server in subsystem i. For a more detailed description of the solution technique see also Tempelmeier and Bürger (2001) or Buzacott, Liu and Shanthikumar (1995).

3.4 Subsystem Analysis

In the subsystem analysis in the iteration the performance parameters and the new throughput is estimated based on the stopped arrival queue approximation introduced by Buzacott Buzacott, Liu and Shanthikumar (1995). It deviates from well known loss systems in that the arrival process is stopped once the buffer to the subsequent server is full. Here, arrival volumes are temporarily adjusted so as to avoid the situation that requests arriving to a full queue are lost. This allows to adjourn arrivals to subsystem i in case buffer i is full for the residual service time of the departure server D_i.

The approximation is based on the assumption of a geometric behavior of the stationary queue length distribution in the $GI/GI/1/N$-queue, i.e.

$$\pi_i = \sigma\pi_{i-1}, \tag{1}$$

and builds on results for the $GI/GI/1/\infty$-queue. The basic idea is to re-scale the results for the $GI/GI/1/\infty$-queue, such that the finite capacity of the $GI/GI/1/N$-queue is accounted for and the maximum number of customers in the system is delimited by $N+1$. Based on the queue length distribution the throughput TH and the expected number of customers N_{cust} in a finite capacity stopped arrival queue can be calculated by

$$TH = \lambda \cdot \frac{1-\rho\sigma^{N-1}}{1-\rho^2\sigma^{N-1}}, \tag{2}$$

where the scaling parameter σ is given by $\sigma = \frac{N-\rho}{N}$ and

$$N_{cust} = \frac{\rho^2(1-c_S^2)}{2(1-\rho)} \cdot \frac{c_A^2+\rho^2 c_S^2}{1+\rho^2 c_S^2} + \rho \tag{3}$$

is an approximation of the average number of customers in the $GI/GI/1/\infty$-queue. For the rather technical derivation of the stopped arrival queue approximation the reader is pointed to the textbook by Buzacott and Shanthikumaer (1993).

Please note that the general procedure of obtaining the throughput from the stationary distribution is not based on the stopped arrival queue approximation, but can be derived from any approximation of the stationary queue length distribution of the $GI/GI/1/N$-queue via

$$TH = \sum_{i>0} \pi_i \mu, \tag{4}$$

where π is the stationary queue length distribution and μ the service rate. This has recently been exploited in Bierbooms, Adan and van Vuuren (2013) and allows for generalization to $G/G/1/N$-queues, i.e. queues with correlated arrival and service processes. In Bierbooms, Adan and van Vuuren (2013), the authors have approximated the arrival and service processes in $G/G/1/N$-service stations by phase-type distributions, such that the subsystems can be solved with standard Markov chain techniques on an extended state space.

While the mean values of the service times of the arrival and departure servers are updated in the forward and backward iteration steps as presented in Algorithm 1, the question remains how to approximate the variation of the arrival and departure processes in the finite capacity queue in the presence of blocking and starvation. The fact that some trains are blocked with probability p_B and receive a prolonged service, whereas others are not, tends to increase the variation of the service process. Hence, for more precise results, an update of the squared coefficient of variation of the arrival and departure servers is given by Buzacott, Liu and Shanthikumar (1995):

$$c_D^2(i) = \frac{(c_S^2(i)+1)S(i)^2 + p_b(i)((c_D^2(i+1)+1)D(i+1)^2 + 2\cdot S(i)D(i+1))}{D(i)^2} - 1.$$

A similar estimate can also be derived for the SCV of the arrival server based on the starvation probability $p_s(i)$:

$$c_A^2(i) = \frac{(c_S^2(i)+1)S(i)^2 + p_s(i-1)((c_A^2(i-1)+1)A(i-1)^2 + 2\cdot S(i)A(i-1))}{A(i)^2} - 1.$$

However, it was found in (Tempelmeier and Bürger, 2001) that Algorithm 1 works best if either c_A or c_D are modified, but not both.

3.5 Performance Indicators

The two main performance indicators we investigate in this paper are the average throughput and the trains' mean running times in the network segments, individually, as well as for the entire train route.

Throughput determination is what the decomposition approach has primarily been developed for. It is obtained as a direct result from the backward forward algorithm. As a performance indicator, the throughput of the queuing network allows to calculate the maximum number of trains that can be accommodated by the system under given service time fluctuations and network correlations. As such, the throughput can serve as an indicator for the absolute achievable capacity, such that the system remains stable. Please note that this does not correspond to the absolute or theoretical capacity used previously in the railway context as the latter is based on deterministic running times. Throughput rather relates to the minimum cycle times in stochastic approaches as in Goverde (2007).

While throughput defines a stability criterion for railway operations on densely operated corridors, most corridors are not operated to capacity as this would imply heavy tolls on running time and delays. This is why the effective running time at a specific traffic load might be better suited as a capacity metric than the maximum throughput. To obtain this performance indicator from the decomposition approach a slight trick is required: We add an additional (virtual) server with infinite waiting space at the beginning of the queuing line that can never be starved and whose service time matches the statistics of train arrivals to the corridor. As this server is never starved, i.e. there always is a train waiting at a service completion epoch, the departure process of this server is identical to the desired arrival process. As a consequence, it produces arrivals to the first "real" server at a the service rate of this virtual server and we can investigate the performance of the corridor given a specific traffic load and statistics.

3.6 Validation

To validate the queuing approach we compare the analytic results to simulation results obtained with a discrete event simulation of rail traffic on the corridor. Extended timetables with several thousand trains are used in the simulation to make the results comparable to the stationary results in the queuing approach and to avoid underestimating system performance parameters due to finite size effects in the simulation. Trains are generated randomly and taken to have random running and stopping times in the simulation. Both arrival intervals and train running and stopping times are taken to be hypo-exponentially distributed, a generalization of the Erlang distribution that allows for matching of continuous values of the SCV. In our model hypo-exponential distributions are matched to the first two moments of train running times on the corresponding infrastructure (cf. Weik and Nießen (2017)).

In Figure 3, the queuing theory approximation of the throughput is compared to the simulation results for different variation of the queue service times at the individual queues in the queuing systems. In addition, the theoretical (absolute) capacity in a deterministic scenario corresponding to the slowest queue's throughput, is depicted as a red dashed line for reference. A line consisting of 20 queues has been considered, where the mean queue service times are chosen randomly uniformly distributed in the interval between 1.5 min and 2.5 min – of the order of minimum headway times in mass transit systems. An average over 20 random system designs, i.e. 20 different infrastructures has been performed.

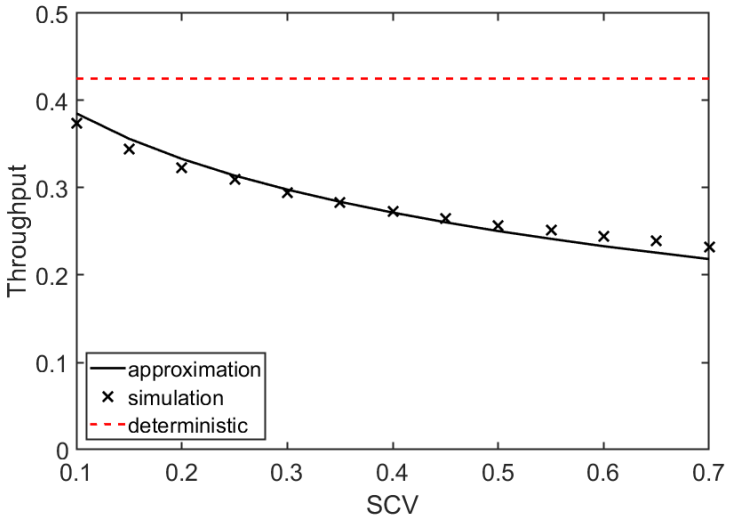

Figure 3: Comparison of queuing approximation and simulation of throughput. Throughput in deterministic systems (reciprocal of processing time of slowest queue) is added as a reference. An average over 20 different queuing infrastructure realizations has been performed.

It can be seen that the queuing method approximates the simulated results reasonably well. Whereas for small variations of service times at the individual queues – which are typical for train running times – the analytic approach tends to slightly overestimate the throughput. For large variation, a slight underestimation of the throughput is found. What is particularly striking in Figure 3 is that the theoretical absolute throughput in a deterministic

system, which is governed by the slowest queue, i.e. the infrastructure segment with the largest running time, largely overestimates the actual throughput of the system, especially for large variations of running times.

Looking at the effective service times of trains at the single queues in Figure 4 for two single realizations with high variation explains why. In a scenario, where all queues have approximately the same service times of about 1 min and $c_S^2 = 0.7$, delays are back-propagated, such that blocking effects run back along the line and yield ever-increasing running times of trains. As a consequence, the entry to the beginning of the line serves as a bottleneck.

To further investigate this effect we consider a second scenario where the first 10 queues have 40% increased mean service times as compared to the second half of the system and the variability of the last 10 queues is significantly lower ($c_S^2(11-20) = 0.1$ vs. $c_S^2(1-10) = 0.7$). A similar scenario could for example occur if the infrastructure state changes on the corridor, such that the mean block length is significantly larger in the first segment of the corridor.

As Figure 4 shows, the effective service times of the last queues remain stable and the back-propagation of delay only occurs within the bottleneck area 1 to 10. These very pronounced spillback effects cannot be covered in a simple deterministic bottleneck analysis, which is why the corresponding throughput in a deterministic scenario is largely overestimated.

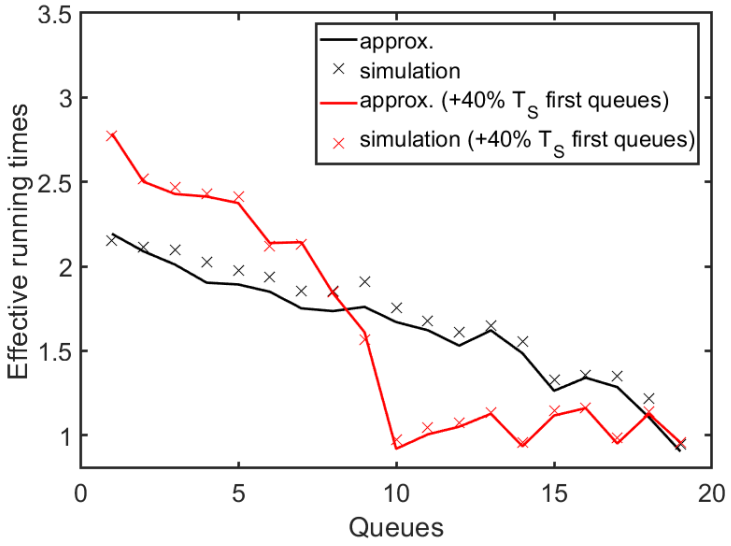

Figure 4: Comparison of approximated and simulated service times at maximum throughput in case of strong variations ($c_S^2 = 0.7$). The black curve depicts a scenario where the effective service times of all queues is in the range of 1 min. In the second (red) scenario, service processes have been updated such that the first 10 queues have 40% higher service times and the second ten queues have reduced variation ($c_S^2(11-20) = 0.1$).

In practice, however, corridors are rarely operated close to saturation, which is why the effective train running times at a specific traffic load are more relevant to estimate service quality – and hence capacity utilization on the corridor. Figure 5 provides a comparison of the queuing approximation and the simulation results for different traffic loads. Again, a system consisting of 20 queues, with railway typical low SCVs of 0.1 is considered. Mean service times are chosen smaller this time to account for shorter block lengths and short stops encountered in metro systems.

As described in the previous Section, a given traffic load is enforced in the queuing model by introducing a virtual starting queue whose service process corresponds to the arrival statistics. It can be seen that the queuing model provides a reliable approximation of the effective service times in this scenario, as well. Only for high load factors over 50% (defined in terms of the deterministic system, i.e. with respect to the slowest queue's service time) the deviations between approximation and simulation start to increase. However, load factors of this order are already close to the theoretical maximum throughput of the system.

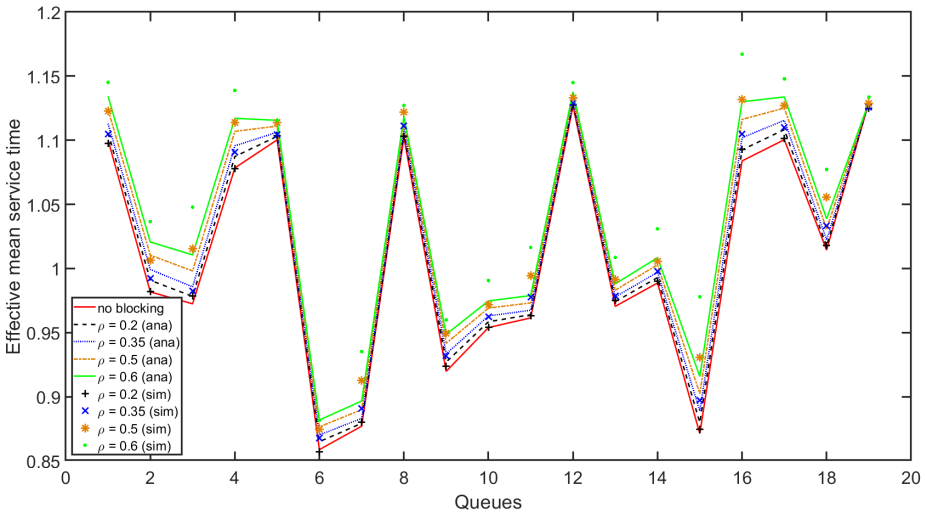

Figure 5: Effective mean train running times on the different track segments as a function of the traffic load on the corridor. Comparison of analytic queuing and simulation results.

4 Case Study: Cologne Commuter Rail Corridor

In the following, we apply the queuing based approach in a case study for performance analysis of Cologne's main commuter rail (S-Bahn) corridor. The central link on this corridor spans the three stops Cologne Hansaring, Cologne Main Station and Cologne Deutz and is delimited by the station Cologne Hansaring to the west and Cologne Posthof intersection to the east (see Figure 6). The central link consists of a total of 11 block sections. It is operated by four commuter rail lines (+ an additional line during peak hours), which all converge on this central link. Each line is operated with a regular frequency of 20 minutes, and additional trains during peak hours. Train services are operated with Bombardier/Alstom trains of type BR 422 and BR 423.

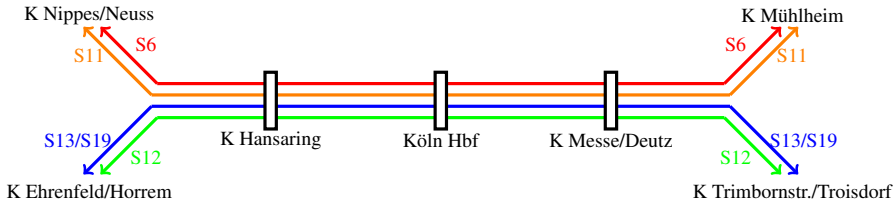

Figure 6: Cologne commuter rail network with central corridor.

The train running times on the corridor are calculated using LUKS®(Janecek and Weymann, 2010), the standard timetable planning tool of DB Netz AG in Germany. The squared coefficient of variation of scheduled stops is assumed to be 0.5, on other block segments 0.1 (see Table 1) to model that stopping times tend to be more variable than free running times on intermediate track segments.

Table 1: Input data Cologne commuter rail corridor. Driving Times in seconds.

Block	1	2	3	4	5	6	7	8	9	10	11
	-	-	stop	-	-	stop	-	-	stop	-	-
S	24	14	103	24	26	92	53	20	77	21	26
c_S^2	0.1	0.1	0.5	0.1	0.1	0.5	0.1	0.1	0.5	0.1	0.1

With the throughput variant of the queuing-based solution routine we find the maximum throughput of the Cologne S-Bahn corridor to be $0.541/min$, which is about 7% less than the deterministic value of the slowest block which has a deterministic throughput of $0.5825/min$, individually. That means that the densest train sequence which still yields stable traffic on the corridor consists of a train frequency of $1/1.85\,min$. The relatively small deviation from the absolute deterministic throughput indicates that, on the whole, variability has a significant, but no dramatic effect on the corridor capacity. The relatively large stopping time fluctuations do not seem to provoke a catastrophic backlog of delays.

However, the effect might be more pronounced on longer corridors with the same parameters. To test this, we prolong the Cologne commuter to generic, longer corridors by adding replicas of itself, i.e. we consider multiples of the corridor. The results are depicted in Table 2. It can be seen that the throughput decreases almost linearly in the system size,

yet with a very moderate slope such that the length of the corridor has a visible, but not dramatic effect on system performance.

Table 2: Throughput as a fct. of corridor length (multiples of Cologne comm. rail corr.).

Corridor length	real	double	3 times	5 times	10 times
Throughput (analytic)	0.541	0.533	0.527	0.519	0.499
Throughput (simulation)	0.496	0.481	0.473	0.465	0.458

In Figure 7, the effective running times on the corridor for varying traffic load are depicted. Like in Figure 5 before, traffic load here refers to the deterministic utilization, i.e. a traffic load of 0.5 corresponds to an arrival rate of $0.5/max(S)$, not to the throughput of the entire queuing network. It can be seen that especially the block segments immediately in front of scheduled stops show a significant increase of running times. Here, trains entering the stations are blocked by preceding trains whose exceeding their scheduled stopping times, such that delays spill back to the previous line segment.

For high traffic load the effective service times of the individual queues approximated with the queuing approach are not as precise as the throughput and tend to underestimate running times as can be seen from Figure 7 and the mean total running times of trains on the entire corridor in Table 3. This deviating behavior can possibly be explained by the calculation of the blocking probability using a two-moment approximation of the queue length distribution in the stopped arrival queue approximation. The approximation quality of the queue correlations and train running times, especially at high traffic loads, could possibly be improved by turning to a more detailed phase-type based modeling of the subsystems in the decomposition approach as, for instance, proposed by Bierbooms, Adan and van Vuuren (2013).

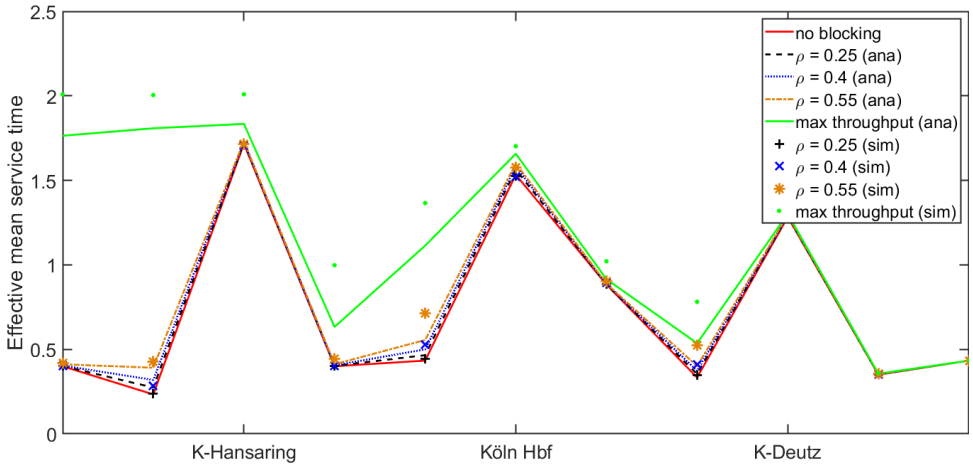

Figure 7: Total running times of trains as a function of traffic load.

Table 3: Running time (min) on corridor as a function of traffic load. The free running time on the corridor is 8 min.

ρ	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
QN	8.18	8.22	8.27	8.32	8.39	8.46	8.54	8.63	8.74	8.87	9.02
sim.	8.10	8.18	8.25	8.38	8.59	8.89	9.08	9.42	10.05	10.79	11.46

5 Outlook to Modeling Options for General Heavy-Rail Corridors

So far, we have discussed applications of single class queuing networks. For metro transit systems with homogeneous traffic concepts this yields realistic results. For more general, heavy rail corridors the situation is more complicated as driving characteristics of trains vary strongly. Different classes of trains need to be considered as running times and stopping policies vary between trains. However, additional complexity is introduced by the possibility that interactions between the train types and shared infrastructure use has to be incorporated. From a purely mathematical point of view this, nonetheless, remains viable as long as the resulting graph is acyclic.

Assuming an optimized dispatching routine that ensures that slow trains are overtaken by fast trains in stations whenever a conflict would occur on the subsequent line segment a generalization to multiple train types can be given. As track occupation conflicts are solved by overtakings and train class changes in stations, queuing centers between stations do not have to correspond to block sections, any more. Instead, a modeling approach similar to Huisman, Boucherie, and van Dijk (2002) can be used, where line segments are represented by three queues: The first and the last one being $G/G/1/0$-queues, enforcing minimum headways on the following line segment and at station entry, and a $G/G/1/\infty$-queue in the middle representing the residual running times on the line segment.

In case of infinite station capacity, this would have implications on slow trains only, which suffer a prolonged station dwell time when put into siding. These overtaking blocking effects can be modeled probabilistically or based on extrinsic parameters from operation data or fast asynchronous simulations of overtakings (also see Lindfeldt (2010) for parameter estimation of overtaking durations). These simulations, which only consider train reordering in stations provide an efficient way to estimate the probability and duration of overtakings in stations (also see Weik et al. (2019)), such that the queuing approach preserves its advantage against parametric simulation. Fast trains would never get delayed as conflicts are assumed to be solved in the previous overtaking station. In reality, however, station overflow, i.e. the situation that a train siding cannot be performed as no station tracks are available, occurs. In this case, a fast train is required to run behind a slower train on the upcoming track segment. We propose to model this situation in the queuing network by a probabilistic class change on the upcoming line segment. The routing is reversed at the next station, where the fast train is routed back to the fast line to get the correct stopping policy.

Figure 8 gives a schematic description for two train classes. The model can of course be extended to multiple trains by introducing additional lines. In this case, additional information on class change probabilities depending on the type of the preceding train is required. Note that the fact that the model assumes class changes occur at the beginning of the line segment is a simplification. The situation that fast trains get caught at some later point on the line segment can be included by splitting the residual running time queue (i.e. the cen-

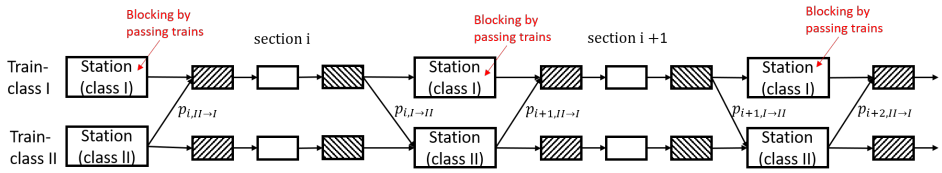

Figure 8: Modeling of heavy rail corridors as a multi-class linear queuing network with customer class switches.

tral $G/G/1/\infty$-queue on line segments) into multiple queues and introducing class change probabilities to the slow class such that the resulting Coxian distribution matches the distribution of running time prolongation for the fast trains.

6 Conclusion

We have investigated the effects of traffic variability such as the stochasticity of train running and stopping times on the capacity of rail corridors. It has been found that irregularities spilling back in the network have a substantial impact on the train frequency that can be stably operated on rail corridors. By adopting queuing theory techniques previously used to describe production lines in manufacturing systems the throughput and the effective train running times on rail corridors without precise timetable data can be obtained. The general validity of the approximation has been demonstrated by comparison to stochastic simulation of train traffic and the method has been applied in a case study for the analysis of the commuter rail's central corridor in Cologne, Germany. Finally, an option to generalize the queuing methodology for general rail corridors allowing to additionally incorporate multiple train classes and blocking/overtaking effects has been discussed.

Acknowledgements
The authors acknowledge the generous support by German Research Foundation (DFG) under Research Grant No. 283085490 "Integral capacity and reliability analysis of guided transport systems based on analytical models" and Research Training Group 2236 "Uncertainty and Randomness in Algorithms, Verification and Logic (UnRAVeL)".

References

Abril, M, Barber, F., Ingolotti, L., Salidom, M. A., Tormos, P., Lova, A., 2007. "An assessment of railway capacity", *Transportation Research Part E: Logistics and Transportation Review*, vol. 44 (5), pp. 774-806, https://doi.org/10.1016/j.tre.2007.04.001.

Arbalete, 2018. "Karte der S-Bahn Köln", modified. https://commons.wikimedia.org/wiki/file:Karte_der_S-Bahn_Köln.png, CC-BY-SA-4.0, last updated: 18-06-15, last accessed: 19-04-02.

Bierbooms, R., Adan, I.J.B.F., and van Vuuren, M., 2013. "Approximate Performance Analysis of Production Lines with Continuous Material Flows and Finite Buffers", *Stochastic Models*, vol. 29 (1), pp. 1–30, http://dx.doi.org/10.1080/15326349.2012.726034.

Büker, T., and Seybold, B., 2012. “Stochastic modelling of delay propagation in large networks”, *Journal of Rail Transport Planning & Management*, vol. 2 (1–2), pp. 34–50, http://dx.doi.org/10.1016/j.jrtpm.2012.10.001.

Burdett, R., and Kozan, E., 2006. “Techniques for absolute capacity determination in railways”, *Transportation Research Part B: Methodological*, vol. 40, pp. 616–632, https://doi.org/10.1016/j.trb.2005.09.004.

Burggraeve, S., and Vansteenwegen, P., 2017. “Robust routing and timetabling in complex railway stations”, *Transportation Research Part B: Methodological*, vol. 101, pp. 228–244, https://doi.org/10.1016/j.trb.2017.04.007.

Buzacott, J. A., Liu, X.-G. and Shanthikumar, J. G., 1992. “Multistage flow line analysis with the stopped arrival queue model”, *IIE Transactions*, vol. 27, pp. 444-455, http://dx.doi.org/10.1080/07408179508936761.

Buzacott, J. A., and Shanthikumar, J. G., 1993. “Stochastic Models of Manufacturing Systems”, Pearson, Englewood Cliffs, US.

Cacchiani, V., Caprara, A., and Toth, P., 2010. “Scheduling extra freight trains on railway networks”, *Transportation Research Part B: Methodological*, vol. 44, pp. 215–231, https://doi.org/10.1016/j.trb.2009.07.007.

Curry, G. L., and Feldman, R. M., 2011. “Manufacturing Systems Modeling and Analysis”, 2nd edition, Springer, Berlin/Heidelberg.

Dingler, M. H., Lai, Y.-C., and Barkan, C. P. L., 2013. “Effect of train-type heterogeneity on single-track heavy haul railway line capacity”, *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 228 (8), pp. 845–856, http://dx.doi.org/10.1177/0954409713496762.

Gershwin, S. B., 1987. “An Efficient Decomposition Method for the Approximate Evaluation of Tandem Queues with Finite Storage Space and Blocking”, *Operations Research*, vol. 35 (2), pp. 291–305, http://dx.doi.org/10.1287/opre.35.2.291.

Goverde, R. M. P., 1998, “The max-plus algebra approach to railway timetable design”, In: *Computers in Railways VI*, pp. 339–350, WIT Press, Southampton, UK.

Goverde, R. M. P., 2007, “Railway timetable stability analysis using max-plus system theory”, *Transportation Research Part B: Methodological*, vol. 41, pp. 179–201, https://doi.org/10.1016/j.trb.2006.02.003.

Hertel, G., 1984. “Exakte Lösung zur Berechnung der Wartegleiszahl vor im Einrichtungsbetrieb befahrenen Streckengleisen bei Nicht-Poisson-Ankünften (G/M/1–Wartesystem)”, *Wiss. Zeitschrift Hochschule für Verkehrswesen*, vol. 31, pp. 195–205.

Huisman, T., and Boucherie, R. J., 2001. “Running times on railway sections with heterogeneous train traffic”, *Transportation Research Part B: Methodological*, vol. 35, pp. 271–292, https://doi.org/10.1016/S0191-2615(99)00051-X.

Huisman, T., Boucherie, R. J., and van Dijk, N. M., 2002. “A solvable queueing network model for railway networks and its validation and applications for the Netherlands”, *European Journal of Operational Research*, vol. 142 (1), pp. 30–51, http://dx.doi.org/10.1016/S0377-2217(01)00269-7.

Janecek, D. and Weymann, F., 2010. “LUKS – Analysis of lines and junctions”, *in Proceedings of the 12th World Congress on Transport Research (WCTR)*, Lisbon, Portugal.

Jensen, L.W., Landex, A., Nielsen, O. A., and Kroon, L.G., 2017. “Strategic assessment of capacity consumption in railway networks: Framework and model”, *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 126–149, https://doi.org/10.1016/j.trc.2016.10.013.

Kerbache, L., and Smith, J. M., 1987. “The generalized expansion method for open finite queueing networks”, *European Journal of Operational Research*, vol. 32, pp. 448–461, https://doi.org/10.1016/S0377-2217(87)80012-7.

Kuckelberg, A., Gröger, T. A., and Wendler, E., 2011. “A UIC-compliant, practically relevant capacity-consumption evaluation algorithm” , *in Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis (RailRome), Rome, Italy*.

Landex, A., Kaas, A. H., Schlittenhelm, B., and Schneider-Tilli, J., 2006. “Evaluation of Railway Capacity”, *in Proceedings of the Annual Transport Conference at Aalborg University*, Aalborg, Denmark.

Landex, A., 2011. “Station Capacity”, *in Proceedings of the 4th International Seminar on Railway Operations Modelling and Analysis (RailRome), 2011*.

Lindfeldt, O., 2010. “Railway operation analysis: Evaluation of quality, infrastructure and timetable on single and double-track lines with analytical models and simulation”, PhD Thesis, KTH Royal Institute of Technology.

Lindner, T., 2011. “Applicability of the analytical UIC Code 406 compression method for evaluating line and station capacity”, *Journal of Rail Transport Planning and Management*, vol. 1 (1), pp. 49–57, https://doi.org/10.1016/j.jrtpm.2011.09.002.

Meester, L. E., and Muns, S. “Stochastic delay propagation in railway networks and phase-type distributions”, *Transportation Research Part B: Methodological*, vol. 41 (2), pp. 218–230, https://10.1016/j.trb.200.602.007.

Mitchell, K., and van de Liefvoort, A., 2003. “Approximation models of feed-forward G/G/1/N queueing networks with correlated arrivals“, *Performance Evaluation*, vol. 52 (2–4), pp. 137–152, https://doi.org/10.1016/S0166-5316(02)00095-0.

Nießen, N., 2013. “Waiting and loss probabilities for route nodes”, *in Proceedings of the 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen), Kopenhagen, Denmark*.

Osorio, C., and Bierlaire, M., 2009. “An analytic finite capacity queueing network model capturing the propagation of congestion and blocking”, *European Journal of Operational Research*, vol. 196, pp. 996–1007, https://doi.org/10.1016/j.ejor.2008.04.035.

Pachl, J., 2014. “Timetable Design Principles” *in Hansen, I. A.; Pachl, J. (eds.): “Railway Timetabling & Operations”*, Chapter 2, pp. 13–46.

Potthoff, G., 1962. “Verkehrsströmungslehre Band 1: Die Zugfolge auf Strecken und in Bahnhöfen” (in German), 1st edition, *Transpress VEB*, Berlin, Germany.

Schwanhäußer, W., 1974. “Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn” (in German), PhD Thesis, RWTH Aachen University.

Schwanhäußer, W., 1994. “The status of German railway operations management in research and practice”, *Transportation Research Part A: Policy and Practice*, vol. 28, pp. 495–500. https://doi.org/10.1016/0965-8564(94)90047-7.

Tempelmeier, H., and Bürger, M., 2001. “Performance evaluation of unbalanced flow lines with general distributed processing times, failures and imperfect production”, *IIE Transactions*, vol. 33, pp. 293–302, https://doi.org/10.1080/07408170108936830.

UIC, “Code 406 – Capacity”, 1st edition, Paris, 2004.

UIC, “Code 406 – Capacity”, 2nd edition, Paris, 2013.

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

Warg, J., 2012. “Effects of increased traffic volume and speed heterogeneity on the capacity of a railway with dense mixed traffic”, In: *Computers in Railways XIII*, pp. 485–497, WIT Press, Southampton, UK, http://dx.doi.org/10.2495/cr120411.

Weik, N., Niebel, N., and Nießen, N., 2016. “Capacity analysis of railway lines in Germany – A rigorous discussion of the queueing based approach”, *Journal of Rail Transport Planning & Management*, vol. 6 (2), pp. 99–115, http://dx.doi.org/10.1016/j.jrtpm.2016.06.001.

Weik, N., and Nießen, N., 2017. “A quasi-birth-and-death-process approach for integrated capacity and reliability modeling of railway systems”, *Journal of Rail Transport Planning & Management*, vol. 7 (3), pp. 114–126, https://doi.org/10.1016/j.jrtpm.2017.06.001.

Weik, N., Warg, J., Johansson, I., Nießen, N., and Bohlin, M., 2019. “Extending UIC 406-based capacity analysis - New approaches for railway nodes and network effects”, *in Proceedings of the 8th International Conference on Railway Operations Modelling and Analysis (RailNorrköping 2019), Norrköping, Sweden*.

Wendler, E., 2007. “The scheduled waiting time on railway lines”, *Transportation Research Part B: Methodological*, vol. 41 (2), pp. 148–158, http://dx.doi.org/10.1016/j.trb.2006.02.009.

Yuan, J., and Hansen, I. A., 2007. ”Optimizing capacity utilization of stations by estimating knock-on train delays”, *Transportation Research Part B: Methodological*, vol. 41, pp. 202–217, https://doi.org/10.1016/j.trb.2006.02.004.