Capacity analysis of railway lines in Germany
- A rigorous discussion of the queueing based approach

Norman Weik*, Nora Niebel*, Nils Nießen*

*Institute of Transport Science, RWTH Aachen University, 52062 Aachen

Abstract

The operation of railway systems requires detailed information on infrastructure capacity. A major challenge, especially in long-term planning, is assessing the quality of operations given very limited information on schedules is available. To this end, analytical models based on a stochastic description of railway systems have found widespread application. We discuss a model for the capacity analysis of railway lines relying on single channel queueing systems. By identifying knock-on delays with waiting times delays can be estimated using methods from stochastics and queueing theory. Mean knock-on delays are used as a quality-dependent indicator of capacity allowing to determine the admissible number of trains for a prescribed level of service.

Though being widely used in Germany the model has not been made fully available to the research community. In this paper two main contributions are made: A new, mathematically rigorous derivation of the pivotal "Strele Formula" for the estimation of knock-on delays, which is based on the convolution of delay distribution functions, is provided. Unlike existing discussions our approach is valid for general independent buffer times. Additionally, we critically review the model assumptions and investigate the "triangular gap problem", an overestimation of capacity resulting from the model’s limitation to pairwise correlations.

Keywords: railway capacity analysis, knock-on delays, delay propagation, queueing, railway infrastructure planning

1. Introduction and Literature Overview

Concise knowledge of the capacity of railway infrastructure is vitally important for the planning, management and operation of railway systems. Depending on the process stage capacity analysis aims to investigate the ability to satisfy the demands of railway operators in scheduling [1] or the stability and robustness of schedules in operations [2]. Input data ranges from fragmentary information about prospective operations to fully constructed schedules. Time criticality enters as a major challenge if dispatching and rescheduling are to be optimized with respect to capacity (cf. [3, 4, 5, 6]). The vastness and complexity of the task – as well as country-specific layout and operations of railway networks – have therefore evoked a multitude of different approaches to determine the feasible number of trains in railway systems.

* This is the Author’s Accepted Manuscript of the following article: N. Weik, N. Niebel, N. Nießen, Capacity analysis of railway lines in Germany – A rigorous discussion of the queueing based approach, Journal of Rail Transport Planning & Management 6(2), pp. 99-115, 2016, which has been published in final form at https://doi.org/10.1016/j.jrtpm.2016.06.001. © 2016. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/
* Corresponding author
Email address: weik@via.rwth-aachen.de (Norman Weik)
URL: http://via.rwth-aachen.de/
Arguably, the most common approach is to use track-occupation as an indicator of capacity usage. An important representative of this class of models is schedule compression documented in UIC Code 406 [7]. It provides a way to determine idle line capacity based on the schedule’s degree of utilization. However, results are hard to compare since track utilization is highly dependent on the structure of the schedule. In addition, the method provides only limited insight into the quality of operations and the schedule’s stability in case of disruptions.

In order to determine capacity in relation to a prescribed level of service that operations are required to meet, different indicators such as punctuality or the magnitude of delays have to be considered. This can be done both from the network manager’s point of view aiming to minimize overall irregularities [8] and from a passenger’s perspective maximizing travel utility [9]. Suggested methods cover microscopic [10, 11, 12] and macroscopic simulations [13] as well as linear and mixed integer optimization [14] aiming to minimize delay-related objective functions [15, 16, 17, 18] or to maximize travel utility [19]. The applicability of both approaches is limited to applications where detailed knowledge about scheduled operations can be used. If input data is scarce and only course information about planned operations is provided a large number of problem instances has to be simulated or solved in order to determine the expected capacity. This usually makes these approaches very time-consuming and hence ineffective for early planning stages.

Analytical stochastic or queueing based approaches are heavily used in this area. In the first case, propagation of delays is determined by convoluting probability distribution functions of delays [13, 20]. In the queueing based approach railway lines and junctions are represented as queueing systems, trains corresponding to requests for the usage of some infrastructure-related service channels (cf. [21, 22] for an overview). Service times correspond to headway times between train runs. While the stochastic nature of these models prohibits the exact individual representation of schedules they are well suited to quickly derive the overall behavior and stability of operations. Moreover, they only require very limited data input: Knowledge of the probability distributions of primary delays, minimum headway times and the projected number of trains and train types is required, but no planned schedule has to be known. In particular, an infrastructure-centered capacity indicator can be derived which allows to compare different system designs regardless of the precise operation concept. Capacity can either be assessed based on the absolute waiting times (delays) [8, 21] or – in addition – by considering their sensitivity w.r.t. the number of trains [23].

We subsequently focus on an analytical queueing based approach to determine the capacity of railway lines and nodes based on the height of knock-on delays. Knock-on delays refer to delays resulting from signalling and safety constraints, dispatchers’ resolution of track occupation conflicts, connection conflicts or fleet rostering conflicts caused by trains operating behind schedule. The presented framework has originally been proposed by Schwanhäußer more than 40 years ago [8], yet software tools based on the approach still build the backbone of capacity analysis in Germany. Although related queueing based approaches have been discussed in the context of timetable quality estimation [1, 24, 25] Schwanhäußer’s pioneering work on the stability of operations has not been made fully accessible beyond the German-speaking region, so far. While Nießen [21] gives a short overview of the approach and the calculation of knock-on delays in general, his focus is on determining a quality measure based on the amount of knock-on delays. An in-depth discussion of the model assumptions and the derivation of the pivotal Strele Formula for the calculation of knock-on delays is missing.
The goal of this paper is to close this gap by presenting a generalized, mathematically rigorous derivation of the Strele Formula. Our approach, which is based on the convolution of delay distribution functions, provides an efficient and elegant way to determine knock-on delays and avoids a time-consuming case by case analysis of delay propagation effects between trains. Instead, it only requires the calculation of certain conditional expectations which can be easily standardized and executed numerically. It also goes beyond previous presentations [8, 26] as it is applicable for general independent buffer times, such that the method can easily be adapted to the needs of specific railway system and timetable characteristics.

We critically discuss modelling assumptions and limitations of the approach. In particular, a lack of exactness on line segments with intersections resulting from the restriction to pairwise correlations between trains and giving rise to the so-called “triangular gap problem” is investigated in Section 3.

2. Method

2.1. Basic Concept

The fundamental characteristic of the method is that it aims to put railway lines to the test individually [8]. Interactions between different railway lines and junctions arising from their embedding in a larger network of lines and junctions are disregarded. While this course of action limits the precision of the analysis it facilitates comparability between different infrastructure layouts and radically reduces computational effort. The original work’s focus was on railway lines [8], yet it has been extended to application in the dimensioning of station threads [27, 28, 29]. In our discussion we focus on the modelling of railway lines. Extensions such as the application to railway junctions are briefly sketched in Section 2.6.

![Figure 1: Schematic representation of queueing systems (cf. [21])]({})

The fundamental concept of the approach is to model railway lines as queueing systems with a single service channel (cf. Figure 1). This assumption does not mean the model’s applicability is restricted to single track railway lines. It extends to lines consisting of two unidirectional tracks common in central European railway systems. In this case, both tracks are treated as individual queueing systems. In the same manner, the method can be extended to three and more tracks.

Originally, the goal was to establish a method allowing to determine the required mean buffer times on railway lines. To this end, a functional relation between knock-on delays and buffer times, is developed [8]. This is achieved by considering the propagation of primary delays due to conflicts between different train runs. Though the model uses a queueing system perspective and is generally referred to as a queueing based approach it is mainly based on mathematical manipulations of delay distribution functions. Results from queueing theory, however, are used to estimate consecutive effects in the propagation of knock-on delays.

Rather than to derive feasible buffer times the model is mainly used to determine the feasible number of trains within a certain time frame. As buffer times are the complement of train service requirements
within a time frame a one-on-one correspondence exists. The overall knock-on delays are used as a quality-
dependent indicator of service quality allowing to establish a functional relation between a prescribed level
of service and the number of trains. We subsequently review the model from this point of view, starting
with a discussion of the basic modelling assumptions in Section 2.2. Section 2.3 is the heart of this paper
where a new, general approximation formula for the average knock-on delay of trains is derived. As an
exemplary application exponential buffer times are inserted and the Strele Formula is recovered in Section
2.4 by performing a specific averaging procedure proposed in [8]. Additionally, the assessment of knock-on
delays in terms of the quality of operations (Level of Service) required for practical application of the method
is discussed in Section 2.5.

2.2. Modelling assumptions

2.2.1. Layout of the queueing system

Apart from the number of channels, queueing systems are characterized by the arrival process, the
service process, the service discipline and possible limitations of the queue size. In the context of this paper
interarrival times and thence the arrival process are defined by minimum headway times, buffer times, and –
if applicable – primary delays of trains. Service times correspond to minimum headway times between trains.

The calculation of minimum headway times is built on the blocking time model (cf. Section 2.2.2) assuming
deterministic travel times and neglecting travel time allowances. The latter are taken to be incorporated
in buffer times [8] such that the multi-objective problem of determining both buffer times and travel time
allowances is avoided.

It is assumed arrival and service times are uncorrelated. This assumption is a potential weak point of
the model given inter-arrival times depend on minimum headway times. It has been remarked in [30], for
example, that this is especially doubtful if periodic timetables are considered. For sufficiently heterogeneous
timetables typical for many German railway lines, however, it is thought that even relatively moderate
primary delays sufficiently disturb timetable structure to justify this assumption. In [8] it is empirically
shown that the variation coefficient of inter-arrival times is close to 1 in this case. It largely deviates from
the statistics of minimum headway times which are found to be approximately constant [8].

The service discipline closely follows German dispatching rules [31] and distinguishes two regimes: Normal
operations and heavily perturbed operations [8]. In normal operations, trains are served according to their
priorities: Long distance passenger services are usually asserted a higher priority than regional passenger
services or freight services. In case of heavily perturbed operations involving the formation of long queues of
waiting trains dispatchers switch to a first-come-first-served strategy (cf. [8]). Train ranks are disregarded
as priority based service would give rise to non-usable idling times further reducing capacity. This complex
service strategy prohibits a direct application of queueing models since priority service, let alone mixed
priority and first-come-first-served disciplines, are extremely hard to analyze mathematically. Therefore,
a combined approach relying on delay distribution functions for the description of normal operations and
queueing models in case of heavily perturbed operations is pursued.

The amount of trains leaving or entering the traffic stream within the line segment represented by the
queueing system is required to be small. This is a necessary assumption since the identification of service
and minimum headway times is problematic given a significant number of trains shares only a minuscule
portion of the route. In this case the line segment has to be divided. Usually, some flexibility is allowed and
minor changes of the train mix (typically below 10% [31]) are accepted.

Finally, the waiting space of the queueing system is assumed to be infinitely large [8]. In case queue
lengths exceed the number of tracks in adjacent stations it is assumed trains can wait on other railway lines
or stations. A reordering of arriving trains is only possible at the beginning of the service process, i.e. at the
boundaries of the investigation area. Overtakings within the line segment, to some extent are accounted for
implicitly as they bear on minimal headway times.

2.2.2. Blocking time model

In queueing system descriptions of railway systems minimum headway times are the central parameter
to describe the service process. They are calculated based on the blocking time model [7, 32] taking into
account microscopic infrastructure characteristics such as signal positions and track clearance detection
beacons. Blocking times correspond to the time interval a line segment is allocated to a single train run
excluding other traffic within this segment. Besides the running times of trains (approach time, driving time
within the block section and clearing time) it also comprises fix velocity-independent time shares (signal
clearing time, release time) depending on the signaling system, as well as the signal viewing time [33]. For a
more detailed discussion of the blocking time’s constituents and the effects of different signaling systems on
blocking times cf. [34]. The totality of a train run thence corresponds to a sequence of blocks, the so-called
“blocking-time stairway” [35] (cf. Figure 2).

Figure 2: Constituents of the blocking time
The minimum headway time between a pair of trains on a railway line is defined as the minimal difference of departure times at the initial station required so their corresponding blocking time stairways do not overlap. In order to create stable schedules buffer times between subsequent train runs have to be supplemented to minimum headway times such that the propagation of (initial) delays between trains can be reduced.

2.2.3. Modelling of buffer times

For heterogeneous timetables scheduled buffer times between train rides are typically assumed to be independently exponentially distributed random variables. This assumption has already been discussed by Potthoff in [36] and has found widespread application in railway engineering to the present day. For dense and highly synchronized schedules in urban transportation, however, a modeling based on fix buffer times seems more appropriate [8].

We subsequently provide a more general discussion of the capacity analysis of railway lines which is completely independent of the distribution of buffer times. It only requires them to be independently distributed as well as to be stochastically independent of service times. The well-known Strele Formula is recovered at the end of our discussion in Section 2.4 by explicitly evaluating the general formulae and inserting exponentially distributed buffer times. By this course of action we ensure a clear documentation of modelling assumptions. It also helps to port the framework to new fields of application involving different schedule characteristics.

2.2.4. Modelling of delays

In the modelling of delays primary delays and knock-on delays are distinguished.

- Primary delays refer to both delays of trains originating beyond the observation space at their point of entry and delays originating within the observation space and not resulting from the resolution of track usage conflicts with other trains.

- Knock-on delays denote delays resulting from the resolution of track occupation conflicts, connection conflicts or rostering conflicts caused by delayed trains in the observation space.

Delays are always assumed to be positive, i.e. trains departing ahead of schedule are not considered, their delays being set to 0. This assumption is motivated by the fact that passenger trains are bound to scheduled departure times in stations and freight trains are only allowed to run ahead of time if they do not obstruct other train rides [8].

Primary delays are modelled by distribution functions of the kind

\[ F_{d,\text{prim}}(t) := \begin{cases} 
0, & t < 0 \\
1 - ce^{-\lambda t}, & t \geq 0,
\end{cases} \]  

with constants \( c, \lambda \). For \( t \geq 0 \) this corresponds to a renormalized exponential distribution. The renormalization factor \( c \) can be interpreted as the probability of the occurrence of primary delays. The use of this distribution in the modelling of delays has been justified both by comparison to empirical data [37, 38, 39] and analytical investigation [40].
In the subsequent discussion different “orders” of knock-on delays are distinguished: First order knock-on delays are knock-on delays resulting from the transmission of primary delays from one train to another. By contrast, higher order knock-on delays refer to knock-on delays due to the transmission of delays between trains which have already suffered knock-on delays in conflicts with other trains [8].

Primary delays, minimum headway times and buffer times are assumed to be stochastically independent in the following.

2.3. Calculation of knock-on delays
2.3.1. Preliminaries

The original derivation of the Strele Formula is based on a graphical case by case analysis of blocking-time stairways in schedules [8]. We subsequently provide a more elegant, mathematically rigorous derivation of a universally applicable approximation formula for the height of knock-on delays, which – upon insertion of specific modelling assumptions such as exponentially distributed buffer times – comprises the Strele Formula. Retaining the same process steps as in [8] we start by discussing first order knock-on delays, the dominant effects in short queues. This corresponds to the normal operation regime, hence requiring the consideration of priority service discipline. Afterwards the heavily perturbed regime where higher-order knock-on delays are prevalent is discussed. In this context, results from queueing theory are applied and an approximate approach to couple the results for first-order and higher order knock-on delays in order to be able to account for the different service disciplines in these two cases is applied.

The calculation of knock-on delays by nature involves correlations between different train runs. In the model only pairwise correlations between different trains are considered; correlations of higher order are not discussed [8]. One of the effects resulting from this simplification is the “triangular gap problem” which will be discussed in detail in Section 3.

Propagation of delays from one train to another depends on the difference of primary delays and buffer times between the two train rides. Primary delays $T_1$, $T_2$, of two trains are independently distributed random variables with parameters $c_1$, $\lambda_1$ and $c_2$, $\lambda_2$ according to Formula (1). Thence the distribution function of the difference of primary delays $\Delta T := T_1 - T_2$ can be determined by convoluting the distribution functions of $T_1$ and $T_2$. Following [13], $F_{\Delta T}$ is expressed as a Stieltjes integral s.t. singular points can be accounted for. A great benefit of this approach is that it reduces the dimensionality of the problem as $T_1$ and $T_2$ no longer have to be considered individually.

Inserting (1) we obtain

$$F_{\Delta T}(t) = \int_{-\infty}^{\infty} F_{T_1}(s + t) dF_{T_2}(s) = \begin{cases} 
1 - c_1 + c_1 \frac{\lambda_1}{\lambda_1 + \lambda_2} c_2 e^{\lambda_2 t}, & t < 0 \\
1 - (1 - c_2 + c_2 \frac{\lambda_2}{\lambda_1 + \lambda_2}) c_1 e^{-\lambda_1 t}, & t \geq 0.
\end{cases}$$

(2)

The distribution function is depicted in Figure 3.
The knowledge about the difference of primary delays and the scheduled buffer times between two train rides is sufficient to determine the transmission of delays from one train to another. Since buffer times and primary delays are stochastically independent, this can again be achieved by convoluting the respective probability distribution functions.

In the following, minimum headway times between two train runs $i$ and $j$ are denoted by $h_{ij}$, buffer times by $b_{ij}$. Mean minimum headway times and buffer times are denoted by $\overline{h}$ and $\overline{b}$, respectively.

### 2.3.2. Delay propagation between succeeding trains

**Formation of knock-on delays:** We subsequently study the mechanisms leading to the propagation of delays from one train to another and hence to the formation of knock-on delays. We first consider two succeeding trains $i$ and $j$ of identical priority.

Trains with identical priority are served according to a first come first served (FCFS) policy. Thus, train $j$ suffers a knock-on delay if and only if the difference of the primary delays is larger than the buffer times, but still does not arrive before train $i$: $b_{ij} < \Delta T \leq b_{ij} + h_{ij}$ (cf. Figure 4b). The FCFS policy implies that train order is switched once $\Delta T > b_{ij} + h_{ij}$. In this case train $i$ receives a knock-on delay if it has to wait for $j$ to clear the track (cf. Figure 4c). If $\Delta T > b_{ij} + h_{ij} + h_{ji}$ the order of trains $i$ and $j$ can be switched without affecting train $i$, s.t. there are no knock-on delays being transferred from $j$ to $i$ (cf. Figure 4d). However, there may be conflicts with other train rides not directly succeeding train $i$ according to schedule. These effects are treated separately in Section 2.3.3.

Note that minimum headway times are not symmetric ($h_{ij} \neq h_{ji}$). Knock-on delays therefore are not continuous in $\Delta T$. In general, a jump is observed once the order of trains is reversed. A schematic representation of the transmitted delay as a function of $\Delta T$ is depicted in Figure 5.

**Priority service:** In railway operations long distance passenger trains generally have a higher priority than regional service trains or freight trains. Thence, in case of conflicting train rides the higher-ranking train is given preferential treatment so as to minimize negative implications to this train.

Let $r(i)$ be the rank of train $i$. In the following, high rank numbers correspond to high priorities, low rank numbers to low priorities. With respect to knock-on delays adoption of a strictly rank-based conflict resolution strategy in operations implies that, in case
Figure 4: Emergence of first-order knock-on delays $K_{12}$ resulting from signalling and safety constraints preventing an overlap of blocking time stairways. The different ways primary delays can be propagated between two trains of equal rank are illustrated. (a) scheduled blocking and buffer times of two succeeding train rides in the absence of delays. (b) $b_{ij} < \Delta T \leq b_{ij} + h_{ij}$ – second (blue) train suffers knock-on delay. (c) $b_{ij} + h_{ij} < \Delta T \leq b_{ij} + h_{ij} + h_{ji}$ – order of trains reversed, first (red) train suffers knock-on delay. (d) $\Delta T > b_{ij} + h_{ij} + h_{ji}$ – no delay propagation between succeeding trains.

Figure 5: Transmitted delay as a function of the difference of primary delays for trains of equal rank. Up to $\Delta T = b_{ij} + h_{ij}$ the succeeding train suffers a knock-on delay. If $\Delta T$ exceeds this value the order of trains is reversed and the first train according to schedule suffers a knock-on delay.
• train $i$ has higher priority than train $j$, train order is only reversed once $\Delta T > b_{ij} + h_{ij} + h_{ji}$. This ensures the higher-ranking train never suffers a knock-on delay due to conflicts with train $j$.

• train $i$ has lower priority than train $j$, train order is reversed once $\Delta T > b_{ij}$.

In order to avoid further complicated case-by-case analyses adding an “operational service time supplement” $t_{d,ij}$ to the service times to account for different train ranks has been proposed [27]:

$$t_{\text{service}} = h_{ij} + t_{d,ij},$$

where $t_{d,ij} = \begin{cases} +h_{ji}, & r(i) > r(j) \\ 0, & r(i) = r(j) \\ -h_{ij}, & r(i) < r(j) \end{cases}$ (3)

$t_{d,ij}$ ensures the higher ranking train never has to wait for the lower-ranking one. The approach has equally been applied in the context of scheduled waiting times where it has been discussed in [1] or [21], for example.

Calculation of knock-on delays: Expressed in a mathematically rigorous and computationally accessible way the expectation value of knock-on delays being transferred between two succeeding trains, subsequently denoted by 1 and 2, is obtained by evaluating the following Stieltjes-Integral:

$$K_{12}^{(1)} = E[K_{12}^{(1)}|b_{12} \leq \Delta T \leq h_{12} + b_{12} + t_{d,12}] + E[K_{12}^{(1)}|h_{12} + b_{12} + t_{d,12} < \Delta T \leq h_{12} + h_{21} + b_{12}]$$

$$= \int_0^{\infty} \left\{ \int_0^{h_{12} + t_{d,12}} s dF_{\Delta T}(s + b_{12}) + \int_{h_{12} + t_{d,12}}^{h_{12} + h_{21}} (h_{12} + h_{21} - s) dF_{\Delta T}(s + b_{12}) \right\} dF_{b_{12}}(b_{12}).$$

After some tedious calculation (see Appendix A.1)

$$K_{12}^{(1)} = B_{12} \int_0^{\infty} e^{-\lambda_2 b_{12}} dF_{b_{12}}(b_{12})$$

with

$$B_{12} := c_1 \lambda_1 \left(1 - c_2 + c_2 \frac{\lambda_2}{\lambda_1 + \lambda_2} \right) \left[ \frac{1}{\lambda_1^2} + \frac{h_{21} - h_{12} - 2t_{d,12}}{\lambda_1} - \frac{2}{\lambda_1^2} \right] e^{-\lambda_1 (h_{12} + t_{d,12})} + \frac{1}{\lambda_1^2} e^{-\lambda_1 (h_{12} + h_{21})}$$

is obtained.

Note that the factor $B_{12}$ only depends on traction parameters (via minimum headway times) and primary delays and does not involve any information about schedule or schedule quality. This information is encoded in the second factor. Depending on the probability distribution of the buffer times the integral may evade analytical treatment requiring numerical calculation.

Also note that the first expectation value corresponds to the knock-on delay transferred to the succeeding train, the second term to those suffered by the preceding train. It can be seen that $t_{d,12}$ ensures one of the integrals vanishes reflecting the fact that the high priority train does not suffer any knock-on delays. In case $r(1) = r(2)$, $t_{d,12} = 0$ and the situation in Figure 5 is recovered.

2.3.3. Heavily delayed trains – Delay propagation between not directly succeeding trains

Up to this point we have only considered delay propagation between trains directly succeeding according to schedule. In case a train is subject to a large primary delay it will not experience conflicts with its direct successor but with subsequent trains instead.
The effective buffer times \( b_{1\rightarrow n} \) between train 1 and train \( n > 2 \) in a sequence of trains corresponds to the sum of minimum headway times \( h_{i,i+1} \) and buffer times \( b_{i,i+1} \) (\( i = 1, \ldots, n-1 \)) between intermediate trains (cf. Figure 6) subtracted by the minimum headway time between train 1 and \( n \). Hence, buffer times consist of a train-specific component \( \tilde{h}_{1n} \) (sum of minimum headway times) and a random component \( \tilde{b}_{1n} \) (buffer times).

\[
b_{1\rightarrow n} = \sum_{i=1}^{n-1} b_{i,i+1} + \sum_{i=1}^{n-1} h_{i,i+1} - h_{1n} = \tilde{b}_{1n} + \tilde{h}_{1n}
\]  

(5)

As in the discussion of the previous section, train 1 suffers a knock-on delay if \( b_{1\rightarrow n} < \Delta T \leq b_{1\rightarrow n} + h_{1n} + t_{d,1n} \), train \( n \) if \( b_{1\rightarrow n} + h_{1n} + t_{d,1n} < \Delta T \leq b_{1\rightarrow n} + h_{1n} + h_{n1} \). Completely analogously to the previous section, the overall knock-on delays are obtained by calculating the expectation value of the transmitted delay between train 1 and \( n \) (see Appendix A.1).

\[
\overline{K}^{(1)}_{1n} = E[K_{1n}^{(1)}|b_{1\rightarrow n} < \Delta T \leq b_{1\rightarrow n} + h_{1n} + t_{d,1n}] + E[K_{1n}^{(1)}|b_{1\rightarrow n} + h_{1n} + t_{d,1n} < \Delta T \leq b_{1\rightarrow n} + h_{1n} + h_{n1}] 
= B_{1n} \cdot e^{-\lambda_{1}\tilde{h}_{1n}} \int_{0}^{\infty} e^{-\lambda_{1}\tilde{b}_{1n}} dF_{\tilde{b}_{1n}}(\tilde{b}_{1n})
\]

By summing over all conflicting trains \( n \geq 2 \) the total first order knock-on delays per train are obtained.

\[
\overline{K}^{(1)} = \sum_{n=2}^{\infty} B_{1n} \cdot e^{-\lambda_{1}\tilde{h}_{1n}} \int_{0}^{\infty} e^{-\lambda_{1}\tilde{b}_{1n}} dF_{\tilde{b}_{1n}}(\tilde{b}_{1n})
\]  

(6)

For general buffer times no closed analytical expression further simplifying Formula (6) can be given. For independently exponentially distributed buffer times \( b_{i,i+1} \), \( \tilde{b}_{1n} \) is Erlang-distributed \( \tilde{b}_{1n} \sim \text{Erl}(\frac{1}{\lambda_{1}},n-1) \) and the integral factorizes. The same holds true for deterministic buffer times.
2.3.4. Higher-order knock-on delays

Apart from the cases studied so far knock-on delays may also form if a train has suffered a knock-on delay and therefore conflicts with other train runs. These so-called "higher-order knock-on delays" are particularly virulent if bunches of trains with small buffer-times are present in the timetable. In this case knock-on delays and "traffic jams" (long queues) can be caused even by relatively small primary delays [8].

In such heavily disturbed operations dispatchers generally switch from a priority based conflict resolution strategy to a first-come-first-served practice. This change reflects that train order is hard to manipulate once a long queue of waiting trains has formed. Moreover, it avoids idling times of the service channel. Thence, as opposed to first-order knock-on delays train priorities are neglected for the calculation of higher order delays [8].

The idea is to identify the overall knock-on delay with the average waiting time in queueing models with FCFS service policy [8]. In a second step, the fraction of first-order knock-on delays within the overall waiting times is estimated. The resulting quotient is then used to scale up the results for first-order knock-on delays given by Formula (6) to cover for all orders of knock-on delays.

For the queueing model an $M/GI/1/\infty$-system is assumed [8]. While the single channel assumption and the infinitely large waiting area have been justified before, the choice of arrival and service process requires some elaboration: If long queues have formed and the timetable is totally perturbed the variation coefficient of the inter-arrival times of trains is close to 1 [8]. This has already been discussed in Section 2.2 to justify the stochastic independence of arrival and service process. Another consequence is that inter-arrival times themselves are uncorrelated such that a Markovian arrival process can be assumed [8]. Schwanhäußer now further assumes this can be transferred to the intermediate regime where the timetable is perturbed, but train rides are not operating entirely independent of schedule. This hypothesis is backed by empirical data showing that if $c_1 = c_2 = 0.52$ and $\bar{t}_{d,prim} = (\bar{\lambda})^{-1} = 21$ min exponentially distributed arrival times can still be assumed [8]. However, the validity of this assumption, especially in case of largely periodic timetables has not been fully investigated.

We denote mean arrival times by $ET_A$, mean service times by $ET_S$, variance of service times by $\text{Var}(T_S)$ and system utilization by $\rho := \frac{ET_S}{ET_A}$. Then, the mean queue length in systems of type $M/GI/1/\infty$ can be calculated using the Pollaczek-Khintchine formula [41, 42]

$$ET_W = \frac{\rho + \lambda \mu \text{Var}(T_S)}{2(\mu - \lambda)}. \tag{7}$$

In order to substitute the share of first order delays in (7) by the results of Sections 2.3.2 and 2.3.3 a necessary precursor is knowing the share of first order knock-on delays in the overall waiting times in Formula (7). Mathematically, this corresponds to the challenging task of determining the average waiting times given only a single customer is waiting for service. To the best of the authors’ knowledge no analytical solution to this problem has been found so far. Therefore, Schwanhäußer proposes estimating this share by considering a special limiting case of Formula (6).

Since the queueing model relies on FCFS service, he only considers trains of the same order. He therefore splits Formula (6) in the two components [8]

$$B_{1n,eq} = C_{1n} \left[ \frac{1}{\lambda_1^2} \left( \frac{h_{1n} - h_{1n}}{\lambda_1} - \frac{2}{\lambda_1^2} \right) e^{-\lambda_1 h_{1n}} + \frac{1}{\lambda_1^2} e^{-\lambda_1 (h_{1n} + h_{1n})} \right] \tag{8}$$
and
\[
B_{1n, \text{diff}} = \frac{C_{1n}}{2} \left[ \frac{h_{n1} + h_{1n}}{\lambda_1} - \left( \frac{h_{n1} + h_{1n}}{\lambda_1} \right) e^{-\lambda_1(h_{1n}+h_{n1})} \right].
\]  
(9)

The first component corresponds to first order knock-on delays resulting from conflicts between trains of equal rank, the second between trains of different rank [8]. For \( B_{1n, \text{diff}} \) it has additionally been assumed that it is equally likely train \( i \) and train \( j \) have higher priority. Thus, both contributions enter in \( B_{1n} \) with equal weight resulting in Formula (9).

Note that buffer times \( b_{i,i+1} \) are independently identically distributed (cf. Section 2.2), such that they are independent of the ranks of trains \( i \) and \( i + 1 \) in particular. Thence, Formula (6) becomes
\[
\bar{K}^{(1)} = \frac{\infty}{\lambda_1} \sum_{n=2} \left( p_{eq} B_{1n,eq} + (1 - p_{eq}) B_{1n, \text{diff}} \right) \cdot e^{-\lambda_1 \tilde{b}_{1n}} \int_0^\infty e^{-\lambda_1 \tilde{b}_{1n}} dF_{\tilde{b}_{1n}}(\tilde{b}_{1n}),
\]  
(10)
where \( p_{eq} \) is the probability of two trains having the same rank.

Schwanhäusser's estimate of first-order knock-on delays in (7) now consists in taking a limit of Formula (10), which mimics service strategy and system behavior in the \( M/GI/1/\infty \)-queueing model. It corresponds to [8]

• setting \( p_{eq} = 1 \),
• assuming trains of equal rank have similar driving parameters, such that \( h_{ij}, h_{ji} \approx \bar{h}_{eq} \),
• taking the limits \( \lambda_i \to 0, c_i \to 1 \) corresponding to high delays and high delay probability.

We denote this limit corresponding highly perturbed arrivals, which are practically independent of schedule, and served using FCFS-service by \( \langle \cdot \rangle \).

The share of first-order knock-on delays in (7) is \( \langle \frac{K^{(1)}_{eq}}{ETW} \rangle \) [8]. This means the quotient of first-order and higher-order delays is
\[
\xi := \frac{\langle K^{(1)}_{eq} \rangle}{1 - \langle K^{(1)}_{eq} \rangle}
\]  
(11)
\( \xi \) is used to estimate the overall knock-on delays on railway lines by scaling up first-order knock-on delays determined in Section 2.3.3 according to the following formula:
\[
K = \bar{K}^{(1)} + \frac{1}{\xi} K^{(1)}_{eq}
\]  
(12)
Note that higher-order knock-on delays are prevalent in heavily perturbed operations. Thence only the equal-rank term in \( \bar{K}^{(1)} \) corresponding to FCFS service discipline is considered in this regime.

Formula (12) is the most general approximation of the average knock-on delays in railway systems with complex regime-change disposition rules. It is applicable to general queueing systems of type \( M/GI/1/\infty \) with only mild assumptions on the service process. Unlike the original derivation given in [8] the discussion above is indifferent of the distribution of buffer times such that it can easily be adapted to meet different system requirements.
2.4. Exponentially distributed buffer times – Strele Formula

As an exemplary application of Formula (12) we subsequently consider the case of exponentially distributed buffer times which will allow us to recover the Strele Formula discussed in [8] and widely used in railway capacity analysis in Germany.

In this case,

\[
\int_0^\infty e^{-\lambda_i b_{ij}} dF_{b_{ij}}(b_{ij}) = \int_0^\infty e^{-\lambda_i b_{ij}} \frac{1}{b_{ij}} e^{-\frac{b_{ij}}{\lambda_i}} db_{ij} = \frac{1}{\lambda_i b_{ij} + 1},
\]

and the first order knock-on delays of directly succeeding trains become (cf. Section 2.3.2)

\[
K^{(1)}_{12} = \frac{B_{12}}{\lambda_1 b_{ij} + 1}.
\]

In case of heavily delayed trains, the effective buffer time \(\tilde{b}_{1n}\) between trains 1 and \(n\) is governed by an \(\text{Erl}(\frac{1}{\lambda_1}, n - 1)\)-distribution as buffer times \(b_{i,i+1}\) between different train rides are considered to be independently identically distributed with parameter \(\frac{1}{\lambda_1}\). \(K^{(1)}_{1n}\) coincides with the moment-generating function of \(\tilde{b}_{1n}\) and we obtain

\[
K^{(1)}_{1n} = B_{1n} e^{-\lambda_1 \tilde{b}_{1n}} \int_0^\infty e^{-\lambda_1 \tilde{b}_{1n}} dF_{\tilde{b}_{1n}}(\tilde{b}_{1n}) = B_{1n} e^{-\lambda_1 \tilde{b}_{1n}} \left(\frac{1}{\lambda_1 \tilde{b}_{1n} + 1}\right)^{n-1}.
\]

As a result, the first order knock-on delays in Formula (10) become

\[
K^{(1)} = \sum_{n=2}^{\infty} \left( p_{eq} B_{1n,eq} + (1 - p_{eq}) B_{1n,diff} \right) e^{-\lambda_1 \tilde{b}_{1n}} \left(\frac{1}{\lambda_1 \tilde{b}_{1n} + 1}\right)^{n-1}.
\]

In order to obtain the Strele Formula additional assertions have been introduced by Schwanhäußer at this point [8]:

- Minimal headway times are replaced by the average minimum headway time. The assumption is justified empirically where a variation coefficient of \(v^2 = 0.113\) is found [8], indicating minimum headway times only exhibit minor fluctuations.

- Train specific delays are replaced by the average delay height (\(\lambda_i = \overline{\lambda} = (\overline{t_{d,prim}})^{-1}\)) and delay probability (\(c_i = \overline{c}\)).

Note that these averaging assumptions erase all train-specific information. Recently, alternative approaches based on Formula (16) retaining train-specific information have been proposed [28, 43, 44]. The topic remains an active area of research.

Exploiting the above assumptions, and following the procedure discussed in the previous section the overall knock-on delay is given by

\[
K = \left(\overline{c} - \frac{\overline{\sigma}^2}{2}\right) \cdot \frac{\overline{t}_{d,prim}^2}{\overline{b} + \overline{t}_{d,prim} \left(1 - e^{-\overline{t}_{d,prim}}\right)} \cdot \left[ p_{eq} \left(1 - e^{-\overline{\tau}_{d,prim}}\right)^2 + (1 - p_{eq}) \frac{\overline{h}_{diff}}{\overline{t}_{d,prim}} \left(1 - e^{-2 \overline{\tau}_{d,prim}}\right) + \frac{\overline{h}}{\overline{b}} \left(1 - e^{-\overline{\tau}_{d,prim}}\right)^2 \right],
\]

the Strele Formula. For the technical details of the derivation of Formula (17) see Appendix A.2. This formula is the backbone of the capacity analysis of railway lines in Germany and has been implemented in various software tools [12, 45].
2.5. Level of service

While Formula (12) and the Strele Formula allow to estimate the amount of knock-on delays based on very limited information the method still lacks an assessment of the amount of knock-on delays in terms of the quality of operations. Therefore, in order to be applicable in capacity analysis a functional relation between knock-on delays and service quality has to be established.

The evaluation function used in Germany is based on a correlation analysis of empirical results of an expert survey [46]. In this survey, railway dispatchers were asked to rate the quality of operations on railway lines in their area of responsibility. The expert rating was then mirrored to the lines’ utilization as well as the knock-on delays. In this study the ”train mix“ was found to be crucial for the quality of operations. More exactly, the share of passenger trains \( p_{pax} \) was identified as the decisive variable in establishing a relation between the height of knock-on delays and service quality [46]. Two main reasons can be given to explain this finding:

- Passenger trains usually have multiple stops within the considered line segment. Thus, they tend to entail heterogeneous operations. This significantly reduces capacity.
- More importantly, delays tend to be less critical for freight services than for passenger services. As a consequence, the acceptable amount of delays is smaller the larger the share of passenger trains.

Based on the survey data a logarithmic correlation between the maximal number of trains which can be operated with satisfactory, market-compliant quality and the share of passenger trains \( p_{pax} \) was found [46]. Inverting the interpolating function the admissible amount of knock-on delay in terms of \( p_{pax} \) within a time-frame of 24 hours is [46]

\[
K_{total, adm.} \approx 370 \cdot e^{-1.3p_{pax}}. \tag{18}
\]

This function is used to define the admissible amount of knock-on delays corresponding to optimal quality of operations. In view of the increase of train velocities and train and velocity heterogeneity in modern timetables the validity of the evaluation function (18) seemed questionable. Yet, in a recent study the results could be largely confirmed. The deviations of the prefactors 370 and 1.3 in Formula (18) were found to be of the order of 7.3% and 3.8%, respectively [44].

Based on the practical (optimal) capacity of railway lines defined by Formula (18) different quality levels can be established (cf. Figure 7). A quality measure is defined by [31]

\[
q := \frac{K_{total}}{K_{total, adm.}}. \tag{19}
\]

Ref. [31] states four different levels of quality on railway lines (cf. Figure 8).

Using this categorization it is possible to determine the admissible knock-on delays corresponding to a specific level of service. By equating the admissible knock-on delay per train and the approximation formula for knock-on delays both the feasible number of trains and the required average buffer times corresponding to a certain Level of Service can be determined.
2.6. Extension to railway nodes

In many railway networks it is found that the limiting factor of capacity are railway nodes rather than railway lines. Nodes generally consist of two main components: the station’s set of tracks and the route node consisting of switches and tracks in the transition region between the railway line and the set of tracks. In analytical capacity investigations of railway nodes in Germany usually the scheduled waiting times \([1, 24]\) and waiting or loss probabilities in GI/GI/n/m-systems are considered \([22, 47]\).

Yet, in case of disruptions, particularly route nodes are prone to becoming the bottleneck of operations in case several trains with overlapping paths enter in quick succession. In order to investigate these effects the method discussed in this paper can be extended to also cover route nodes. To this end, route nodes are decomposed into so-called sectional route nodes (SRN), sections which can exclusively be used by one train run only, in the sense that any train entering this section excludes all other traffic within this segment \([26, 27]\). Knock-on delays developing in these single-channel queueing systems can then be calculated in the same manner as on railway lines.

In order to investigate the entire network of SRNs building the route node, the latter is interpreted as a single channel queueing system with reduced service times where the reduction depends on the number and frequency of mutually non-exclusive train routes in the route node. For a detailed discussion of this field the reader is referred to the literature \([26, 28, 29]\).

3. Model limits – Triangular gap problem

In 1978 already, Schwanhäußer mentioned potential shortcomings of the model due to its limitation to pairwise correlations between trains \([27]\). We subsequently discuss and provide a solution to a specific
manifestation of this model property, which leads to an over-estimation of capacity. It is observed in capacity analysis of line segments with intersections where three (or more) succeeding trains only partially share a common route.

Figure 9 depicts the blocking time stairways of a triple of trains. Even though trains $i$ and $k$ follow each other with minimum headway times, the blocking time stairway forms a gap in which a third train $j$ – only sharing a short section of the investigation area - can run without any conflicts. We subsequently refer to this phenomenon, introduced by Schwanhäußer in [27], as "triangular gap problem (TGP)". It occurs if the minimum headway time between trains $i$ and $k$ is bigger than the sum of minimum headway time between train $i$ and $j$ and between train $j$ and $k$:

$$h_{ik} > h_{ij} + h_{jk}.$$  \hspace{1cm} (20)

Figure 9: Instance of the triangular gap problem. Neglecting correlations between three and more trains leads to underestimated service times if $h_{ij} + h_{jk} < h_{ik}$.

As the model is based on interaction between two trains, the influence of small minimum headway times within the triple of trains is overrated - thus the average minimum headway time is underestimated, leading to a higher calculated capacity. This can only be considered precisely when expanding the model to interactions between three or more trains.

A sketch of an approximate solution was recently established by manipulating the minimum headway times based on the probability of occurrence of the TGP [48]. The minimum headway time of the train located in the gap is extended to satisfy the following equation: $h^*_{ij} + h^*_{jk} = h_{ik}$. Hereby, the sum of minimum headway times in chronological order of three regarded trains equals the minimum headway time between the two outer trains. The increase of minimum headway times $h_{ij}$ and $h_{jk}$ to $h^*_{ij}$ and $h^*_{jk}$ is done equidistantly [48], i.e. both are attributed half the distance

$$\Delta h_{ijk} = h_{ik} -(h_{ij} + h_{jk}).$$  \hspace{1cm} (21)

The adjusted minimum headway times $h^*$ are only applicable to those trains within the TGP. Therefore minimum headway times are increased proportional to the probability of the occurrence of a train $j$ within
a gap between trains \( i \) and \( k \). The supplement is calculated by

\[
h_{add,ij} = \frac{p_{TGP}}{p_{i-j|j}} \cdot \frac{\Delta h_{ijk}}{2}
\]  

(22)

with

- \( p_{TGP} \): probability of TGP. \( p_{TGP} \) can be calculated both based on the frequencies of trains or based on an existing schedule.

- \( p_{i-j|j} \): Ratio of trains on route \( j \) running behind a train on route \( i \) w.r.t. all trains \( j \).

Adding the supplements to the corresponding minimum headway times represents the usage of the gap by another train in the capacity calculation. The modified headway times serve as input parameter into the Strele Formula.

**Example:** As an example, we assume minimum headway times between trains continuing on the main line in Figure 9 is 6 min, minimum headway times between exiting (red) trains and continuing trains is only 2 min and between two exiting (red) trains 4 min. Furthermore, the frequency of all train routes (exiting and continuing) is assumed to be equal.

In this case

\[
z_{ik} = 6 \text{ min} \geq 2 \text{ min} + 2 \text{ min} = z_{ij} + z_{jk},
\]

where trains \( i \) and \( k \) continue on the main line and train \( j \) exits at the intersection. The probability of this train sequence and hence the TGP is

\[
p_{TGP} = p_{ij} \cdot p_{jk} = 0.5 \cdot 0.5 = 0.25.
\]

Evaluating the supplements for the minimum headway time one gets

\[
h_{add,ij} = \frac{p_{TGP}}{p_{i-j|j}} \cdot \frac{\Delta h_{ijk}}{2} = \frac{0.25}{0.5} \cdot \frac{6 \text{ min} - 2 \text{ min} - 2 \text{ min}}{2} = 0.5 \text{ min}
\]

Thence the average minimum headway time increases from \( \bar{h} = \frac{6 \text{ min} + 2 \text{ min} + 2 \text{ min} + 4 \text{ min}}{4} = 3.5 \text{ min} \) to \( \bar{h} = \frac{6 \text{ min} + (2+0.5) \text{ min} + (2+0.5) \text{ min} + 4 \text{ min}}{4} = 3.75 \text{ min} \), which is an increase of about 7\%, such that the triangular gap problem can indeed lead to a significant underestimation of knock-on delays.

4. Outlook and future research

The Strele Formula has proved a reliable and effective tool in railway capacity analysis in Germany. Its queueing system basis and the stochastic description of delay propagation and accumulation requires only limited information on operations and schedule structure and makes the approach particularly suitable in long-term planning and capacity prognosis.

We have provided a rigorous discussion of the underlying model, its modelling assumptions as well as its limitations. By introducing convolutions of delay distribution functions a significant simplification of the computation as well as a generalization to general independent buffer times was achieved. This is particularly
relevant for possible adaptations to different areas of application, possibly involving a different dispatching strategy or schedule structure.

Apart from the discussion of the modelling approach in general and the Strele Formula and its application in capacity analysis in Germany in particular we have also pinpointed shortcomings due to the underlying assumptions. The development towards more and more periodic timetables in Europe, for instance, puts the assumption of exponentially distributed uncorrelated inter-arrival times as well as the independence of arrival and service processes in question. Delays can certainly be expected to destroy correlations to a certain extent. In [49] it is argued a mixture of several periodic timetables can be modelled by an exponential distribution. Alternatively, the usage of fix (deterministic) buffer times has been proposed in the context of very dense and homogeneous timetables in urban transport [8]. Yet, the applicability of the model for highly periodic timetables has not conclusively been investigated, so far.

From a mathematical point of view the identification of first-order knock-on delays in the overall queueing times is particularly unsatisfactory. Finding a more rigorous way to identify first-order knock-on delays in the overall waiting times poses an interesting and very challenging problem which could lead to a significant improvement of the model.

Apart from that, Schwanhäußer’s averaging assumptions [8] have been a precursor of debates. The distinction between trains of equal and different priority is quite restrictive. Keeping the train specific information as well as generalizing to an $M/GI/1/\infty$-queueing model in the calculation of higher-order knock-on delays should be possible without too much effort. A prototypical evaluation of this more detailed description of service process and train characteristics is given in [43, 44] and for a different layout of the queueing system in [28]. Still, adapting the framework and the Strele Formula to train-specific knock-on delays remains an active field of research.

References


Appendix A. Calculus

Appendix A.1. First order knock-on delays

In this paragraph first order knock-on delays $\overline{K}^{(1)}_{12}$ are calculated.

\[
\overline{K}^{(1)}_{12} = E[\overline{K}^{(1)}_{12} | b_{12} < \Delta T \leq b_{12} + h_{12} + t_{d,12}] + E[\overline{K}^{(1)}_{12} | b_{12} + h_{12} + t_{d,12} < \Delta T \leq b_{12} + h_{12} + h_{21}]
\]

\[
= \int_0^\infty \left\{ \int_0^{h_{12} + t_{d,12}} e^{-\lambda_1 s} ds + \int_0^{h_{12} + h_{21}} (h_{12} + h_{21} - s) e^{-\lambda_1 s} ds \right\} dF_{b_{12}}(b_{12})
\]

\[
= c_1 \lambda_1 \left( 1 - c_2 + \frac{\lambda_2}{\lambda_1 + \lambda_2} \right) \cdot \int_0^\infty e^{-\lambda_1 b_{12}} \left\{ \int_0^{h_{12} + t_{d,12}} e^{-\lambda_1 s} ds + \int_0^{h_{12} + h_{21}} (h_{12} + h_{21} - s) e^{-\lambda_1 s} ds \right\} dF_{b_{12}}(b_{12})
\]

\[
= C_{12} \left[ \frac{1}{\lambda_1^2} + \left( \frac{h_{21} - h_{12} - 2t_{d,12}}{\lambda_1} - \frac{2}{\lambda_1^2} \right) e^{-\lambda_1 (h_{12} + t_{d,12})} + \frac{1}{\lambda_1^2} e^{-\lambda_1 (h_{12} + h_{21})} \right] \cdot \int_0^\infty e^{-\lambda_1 b_{12}} dF_{b_{12}}(b_{12})
\]

The third line results from the insertion of the the distribution function $F_{\Delta T}$. As delays can only be propagated if the difference of the primary delays of two succeeding trains is positive only the branch for $t \geq 0$ has to be considered. As $F_{\Delta T}$ is continuously differentiable within this branch the substitution

\[
\int g(s) dF_{\Delta T}(s) = \int g(s) F'_{\Delta T}(s) ds,
\]

fore some function $g$ can be made, which gives rise to the prefactor $C_{12}$.

Delay propagation between trains not directly succeeding according to schedule relevant for heavily delayed trains is calculated completely analogously. In this case, $h_{12}$, $b_{12}$, $t_{d,12}$ are substituted by their counterparts $h_{1n}$, $b_{1n}$, $t_{d,1n}$ and an additional factor $e^{-h_{1n}}$ resulting from splitting $b_{1 \rightarrow n}$ into $b_{1n} + \tilde{h}_{1n}$ in $F_{\Delta T}$ enters.

Appendix A.2. Recovery of the Strele Formula

In this paragraph the Strele Formula is recovered starting from Formula (16) by applying Schwanhäußer’s averaging technique [8] discussed in Section 2.4.

\[
\overline{K}^{(1)} = \sum_{n=2}^{\infty} \left( p_{eq} B_{1n,eq} + (1 - p_{eq}) B_{1n,disf} \right) e^{-\lambda_1 \tilde{h}_{1n}} \left( \frac{1}{\lambda_1 \tilde{b} + 1} \right)^{n-1}.
\]  

(A.1)

Inserting the averaging assumptions Formula (16) significantly simplifies to

\[
\overline{K}^{(1)} = \left( p_{eq} B_{12,eq} + (1 - p_{eq}) B_{12,disf} \right) e^{-\frac{\lambda_1 \tilde{h}}{\lambda_1 \tilde{b} + 1}} \sum_{n=2}^{\infty} e^{-\frac{\lambda_2}{\lambda_1 \tilde{b} + 1}} \left( \frac{1}{\lambda_1 \tilde{b} + 1} \right)^{n-2}
\]

(A.2)
with
\[ B_{12,eq} = \frac{c}{\lambda} \left( 1 - \frac{c^2}{2} \right) (1 - e^{-\lambda h_{eq}})^2 \quad \text{and} \quad B_{12,\text{diff}} = \frac{c}{\lambda} \left( 1 - \frac{c^2}{2} \right) h_{\text{diff}} (1 - e^{-2\lambda h_{\text{diff}}}). \]

Evaluating (A.2) we obtain
\[ K^{(1)} = \frac{\tau - \frac{c^2}{\lambda^2}}{\lambda} p_{eq} (1 - e^{-\lambda h_{eq}})^2 + \frac{(1 - p_{eq}) \lambda h_{\text{diff}} (1 - e^{-2\lambda h_{\text{diff}}})}{\lambda b + 1 - e^{-\lambda b}}. \tag{A.3} \]

We now turn to the evaluation of higher order delays. Since minimum headway times are assumed to be constant the underlying queueing system further simplifies to \( M/D/1/\infty \). With \( ET_S = \frac{1}{h} \) and \( ET_A = \frac{1}{h + b} \) (primary delays, on average, do not change arrival intervals) the utilization becomes \( \rho = \frac{h}{h + b} \).

Pollaczek-Khintchine then entails
\[ ET_W = \frac{\rho \cdot ET_S}{2(1 - \rho)} = \frac{h^2}{2(h + b)} \tag{A.4} \]
for the average waiting time.

Finally, the limit \( \langle K_{eq}^{(1)} \rangle \) is determined (cf. Section 2.3.4), where the assumption of equal service times allows to identify \( h_{eq} = \bar{h} \).

\[ \langle K_{eq}^{(1)} \rangle = \frac{\bar{h}}{h + b} \cdot ET_W, \quad \text{thus} \quad \xi = \frac{\bar{h}}{h} \quad \text{(cf. Formula (11)).} \tag{A.6} \]

Inserting (A.6) in (12) the overall knock on delay is given by
\[ K = \frac{\tau - \frac{c^2}{\lambda^2}}{\lambda} \cdot \frac{p_{eq} (1 - e^{-\lambda h_{eq}})^2 + \frac{(1 - p_{eq}) \lambda h_{\text{diff}} (1 - e^{-2\lambda h_{\text{diff}}})}{\lambda b + 1 - e^{-\lambda b}}}{\frac{1}{h} + \frac{1}{b}} \tag{A.7} \]
the Strele Formula.

Rewriting \( \lambda = \frac{1}{t_{d,\text{prim}}} \) the more familiar form
\[ K = \left( \tau - \frac{c^2}{2} \right) \cdot \frac{\frac{\tau^2}{t_{d,\text{prim}}}}{b + \frac{\tau^2}{t_{d,\text{prim}}}} \cdot \left[ \frac{p_{eq} (1 - e^{-\lambda h_{eq}})^2 + \frac{(1 - p_{eq}) \lambda h_{\text{diff}} (1 - e^{-2\lambda h_{\text{diff}}})}{\lambda b + 1 - e^{-\lambda b}}}{\frac{1}{b} + \frac{1}{b}} \right] \tag{A.8} \]
is obtained.