Microscopic railway capacity assessment of heterogeneous traffic under real-life operational conditions

Journal Article

Author(s):
Lordieck, Jan; Nold, Michael; Corman, Francesco

Publication date:
2024-06

Permanent link:
https://doi.org/10.3929/ethz-b-000667887

Rights / license:
Creative Commons Attribution 4.0 International

Originally published in:

Funding acknowledgement:
181210 - DADA - Dynamic data driven Approaches for stochastic Delay propagation Avoidance in railways (SNF)
Microscopic railway capacity assessment of heterogeneous traffic under real-life operational conditions

Jan Lordieck \textsuperscript{a, b,}\textsuperscript{*}, Michael Nold \textsuperscript{a}, Francesco Corman \textsuperscript{a}

\textsuperscript{a} Institute for Transport Planning and Systems, ETH Zürich, Stefano-Franscini-Platz 5, 8093, Zürich, Switzerland
\textsuperscript{b} Rapp AG, Switzerland

ARTICLE INFO

Keywords:
Railway capacity
Delay analysis
Heterogenous traffic
Trajectory reconstruction
Open data

ABSTRACT

A key strategy for increasing railway capacity is utilizing infrastructure more efficiently. While much research has been completed on methods for assessing railway capacity, very little has focused on the details of capacity utilisation, such as assessing the various ways trains use capacity, the impacts of specific blocking time components, and how train dynamics (accelerating, cruising, braking, and dwelling) affect capacity.

This paper presents a methodology for comparing planned occupancy to actual occupancy under real operations and applies it in a case study. The methodology is based on identifying a critical path which represents an extension of bottleneck concept presented in UIC leaflet 406. The methodology was applied in a case study to determine the specific blocking times and train dynamics which cause a blocking time gap for a sequence of trains, both a-priori and a-posteriori, after considering the operational variations. The analysis of real operations with variations in train trajectories shows that capacity occupation is mostly influenced by train sequence heterogeneity in the original schedule. The varying effects of operations have a smaller but relevant impact. The methods developed in this paper can be used to help assess railway capacity under real operations.

1. Introduction

Railway traffic is growing worldwide due to its space, energy, and cost efficiency as well as reduced greenhouse gas emissions compared to private motorised transportation. Given the high cost and long construction times for railway infrastructure (Khadem-Sameni et al., 2010a; Wang et al., 2020), most railways attempt to accommodate this growth by running more trains on their existing networks. However, many highly used rail corridors already have insufficient capacity (Nold and Corman, 2023; Hartjes, 2019; Stalder, 2020). While building new infrastructure is a long-term solution for increasing railway capacity, in the short-term railways must use the capacity of existing infrastructure as efficiently as possible.

Although there are many definitions and assessment methods for railway capacity, it is one of the most often used targets in policy making (Rotoli et al., 2016). Today, capacity is typically defined as the maximum number of trains that can be operated on a specific railway infrastructure in a specific time interval under given operational conditions (Abril et al., 2008; Jensen et al., 2020; Weik and Niejen, 2020). There are also many indicators available for describing capacity (e.g. capacity consumption versus occupation, ratios versus times, etc.).
In this paper, We define capacity occupation as the sum of the time durations of all train paths in a specified timeframe which block the infrastructure (i.e., the time a train is exclusively occupying infrastructure in a block section, also known as occupancy time in UIC norm 406) (Besinović and Goverde, 2018). The difference between the time of capacity occupation by train operation and the total timeframe yields the available capacity as a time value, which can then be transformed into a number of trains using various different methods. If the available capacity is greater than zero, the specified train timetable can be potentially executed. However, there is often a gap between planned and realized capacity occupation. Therefore, it is important to better understand which factors influence capacity occupation (Besinović and Goverde, 2018) and to consider realistic operational conditions when planning timetable adjustments and infrastructure investments.

This research examines how train dynamics (acceleration, cruising, braking, and dwelling in stations) affects capacity usage. These train dynamics depend on rolling stock characteristics (e.g., acceleration and braking performance), the infrastructure (e.g., block layout, signalling system, speed restrictions), and the operation (e.g., train mix and sequence, train priorities) (Besinović and Goverde, 2018). While train dynamics are implicitly included in existing capacity assessment methods, these methods rarely consider deviations in train dynamics from planned operations or day-to-day variations. Both of these greatly affect capacity occupation but have not been fully researched (Khadem Sameni and Moradi, 2022).

Therefore, the underlying research questions are: How much capacity is occupied by which blocking time component and train dynamic in planned and in realized operation? How does the planned capacity occupation differ from realized capacity occupation? and How can the deviation in capacity occupation be analysed?

Understanding how capacity has actually been used and how occupation changed because of operational variations helps planners to improve railway operation. It enables for example (1) to tune timetables by adjusting the running time reserves according to measured operational deviations to enable highest possible efficiency while still considering disturbances where necessary, (2) to identify which specific aspects of the operation occupy most capacity and thereby informing infrastructure adjustment planning and rolling stock acquisition tailored to specific needs or (3) to better understand which operational precision is necessary to reach the desired capacity and operational stability, hence improving the interplay between planning and operation. (Landex, 2008).

This research proposes a methodology for reconstructing train trajectories in real-life operations, and then extending the standard compression method for estimating capacity (UIC norm 406) to analyse capacity in realized operations. This methodology is applied to a case study on the Zurich-Chur line in Switzerland. The results show a clear difference between planned and real capacity occupation, as well as the impact of train sequence heterogeneity and variations in operation. The case study shows how the methodology could be helpful for improving many aspects of railway planning including infrastructure decision-making and evaluation of schedule robustness.

The paper continues as follows. Chapter 2 summarises the literature on railway capacity from the point of view of influencing factors, and analysis of operations. Chapter 3 describes the general approach. Chapter 4 describes implementation of the methodology. Chapter 5 summarises the case study and Chapter 6 presents an analysis showing how the methodology can be used to improve railway planning using the case study as an example. Chapter 7 presents conclusions.

2. Literature review

2.1. Capacity assessment methods

Railway capacity assessment methods can be grouped into four categories as summarised in Table 1. Some authors merge categories, for example, optimization and analytical methods (Abril et al., 2008), or do not consider simulation techniques as their own category (Khadem-Sameni et al., 2010b). Furthermore, many approaches combine different methods, prominently timetable compression, optimization, and simulation techniques (Jensen et al., 2020; Pouryousef et al., 2015; Weik et al., 2016, 2020). This paper focuses on the capacity assessment of links, (i.e. line infrastructure); similar approaches are available for nodes, (i.e. station, junction, and interlocking infrastructure).

All methods have specific advantages and disadvantages which can be overcome by combining them, however, this increases complexity. All presented methods are applied in a European context, except parametric models, which are developed in the North American context of railways which do not follow a strict schedule (Besinović and Goverde, 2018; Pouryousef et al., 2015). Cross-application of European and American methods has been suggested (Pouryousef et al., 2015), but has not yet been used (Besinović and Goverde, 2018). The following sections review methods most appropriate for this research; for a broader review of methods see for example Khadem Sameni and Moradi (2022).

2.2. Capacity assessment methods without requiring a schedule

Several methods exist to assess the capacity of a railway infrastructure without requiring a schedule. In simple cases, capacity can be calculated using an empirical formula which utilizes the minimum headway in the bottleneck section for double-track lines or the distance between points allowing for crossing for single-track lines (Abril et al., 2008). However, identifying the bottleneck on larger sections or for entire networks is non-trivial, and the use of mixed train sets with multiple possible sequences complicates the problem

---

1 To avoid confusion, we only refer to the term “timetable” in relation to the method of timetable compression in this article. Otherwise we use “schedule”.

---

2 Journal of Rail Transport Planning & Management 30 (2024) 100446
Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Methodological Foundation</th>
<th>Typical application</th>
<th>Example publication for method development and application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analytical</td>
<td>Analytical formulae based on minimum headway</td>
<td>Simple lines and nodes</td>
<td>Cerny et al. (2018); UIC, 1996; Widyastuti and Budhi (2020)</td>
</tr>
<tr>
<td>Timetable compression</td>
<td>Timetable lines, nodes, and complete networks</td>
<td>Complex lines, nodes, and complete networks</td>
<td>Meirich (2017); Weik et al., 2016</td>
</tr>
<tr>
<td>Queuing theory</td>
<td>Queuing theory</td>
<td>Capacity assessment based on queuing theory with a defined quality parameter and probabilistic train sequences without a timetable</td>
<td>Meirich (2017); Weik et al., 2016</td>
</tr>
<tr>
<td>Parametric</td>
<td>Parametric statistics (Regression, neural networks, etc.)</td>
<td>Simple lines</td>
<td>Lai et al. (2014); Lai and Barkan (2009)</td>
</tr>
<tr>
<td>Optimization</td>
<td>Optimization mathematical optimization, graph theory</td>
<td>Complex lines, nodes, and complete networks</td>
<td>Burdett (2015b); Azadi Moghaddam Arani et al., 2019; Jensen et al., 2020</td>
</tr>
<tr>
<td>Simulation</td>
<td>Simulation various simulation techniques (discrete/continuous, synchronous/asynchronous, etc.) and software tools (Opentrack, Railsys, etc.)</td>
<td>Complex lines, nodes, and complete networks</td>
<td>Goverde et al. (2013); Weik and Niejen (2020)</td>
</tr>
</tbody>
</table>

* The examples are neither exhaustive nor do they sometimes fit perfectly into one category, however, they represent a typical application and show the methods’ capabilities.

further (Burdett and Kozan, 2006). Capacity assessment without knowing the schedule is normally used for determining the capacity of new infrastructure when a schedule is not yet available, or when assessing an uncertain future demand. Jensen et al. (2017) call this strategic capacity assessment and note that traditional methods based on timetable compression are not capable of solving this problem. Multiple approaches exist.

A widely applied approach in Germany is based on single-channel queuing systems (Weik et al., 2016). Based on a probability for each train category following each other and a predefined minimum headway for every combination of trains, the mean minimum headway can be calculated for a specific line section (Meirich, 2017; Weik et al., 2016). Using the Strele-Formula developed by Schwanhäußer (1973) the mean waiting time considering unplanned delays can then be calculated and capacity be assessed (Meirich, 2017; Weik et al., 2016). The application requires either good knowledge of the investigated line section and its operation or making multiple non-trivial assumptions for the minimum headway of each train category combination, probability of initial delays, and mean initial delay.

De Kort et al. (2003) study the effect of single or double-track tunnel sections on the capacity of a proposed (and by now realized) high-speed line in the Netherlands utilizing a max,+ algebra optimization approach. Over several years Burdett et al. worked on an optimization algorithm to assess the capacity of railway lines without a schedule, considering also delays and further real-life operation characteristics (Burdett, 2015a, 2015b; Burdett and Kozan, 2006). Recently, approaches that assess an infrastructure’s capacity more realistically have been proposed. Jensen et al. (2017) develop a method in which all possible permutations of a sequence of a given set of trains are tested on their feasible operation for a given infrastructure. The method utilizes a simplified infrastructure graph for which a minimum headway graph for each train set is computed. The longest path in this headway graph marks the time needed to run a sequence of trains. The output is a cumulative distribution of capacity occupations which describes the share of sequences that can be run on the infrastructure. Jensen et al. (2020) apply this approach when developing an optimization algorithm that produces passenger-friendly train sequences to determine the feasibility of a realistic future train sequence. Liao et al. (2021) can determine the schedule that allows for the maximum amount of train services including also the limitations due to vehicle usage. The effects of heterogeneous services are identified.

These approaches have in common that they simplify infrastructure and operations. They consider infrastructure constraints, blocking times, and further operational characteristics implicitly. The output is nevertheless an accurate value for capacity. The main advantage of optimization approaches compared to simple empirical formulas is obtaining an optimally saturated schedule (Wang et al., 2020) and thereby a reliable value for capacity or even a distribution of possible future sequences of trains. Capacity assessment with queuing models as utilized in Germany also includes stochastic elements accounting for different train sequences (Weik et al., 2016). However, the results do not show where and by which train capacity is occupied as well as which factors contribute to capacity occupation.

2.3. Capacity assessment methods given a schedule

Multiple methods to assess the capacity given a planned schedule exist. The methods can be categorized into analytical and
simulation approaches; optimization and parametric approaches only play a minor role when a schedule is present (see for example Abril et al., 2008; Bešinović and Goverde, 2018). The planned schedule results in a given supply, and associated heterogeneity in train sequence. The most widely applied method in Europe, with variations though, is the timetable compression method presented in the UIC leaflet 406 (Pouryousef et al., 2015; UIC, 2004, 2013; Weik et al., 2020).

It is based on blocking time theory. Simply speaking, the blocking time diagram of a line section is compressed reducing all gaps in a blocking time stairway to zero until blocking time stairways of all pairs of consecutive trains touch each other in at least one block. The time between the start of the occupancy of the first train in the investigated interval until the end of the occupancy of the last train inside the same interval represents the capacity occupation time in a block section. Adding static buffer times yields the capacity consumption. The capacity of the investigated line section is measured in the block section wherein the capacity occupation or consumption time is maximal. This block section is often called bottleneck block (Abril et al., 2008). To determine the capacity of an infrastructure with this method beyond the present train sequence, the compressed timetable is enriched with additional duplicates of trains already present in the schedule until the maximum occupation or consumption time reach the time span of the investigated interval (UIC, 2004, 2013).

The first edition of the norm (UIC, 2004) has been highly criticized, especially for the decomposition of the infrastructure into very small segments which overestimates the capacity of connected infrastructures (Landex et al., 2008). Further points of criticism are its inapplicability in stations and switch areas as well as the highly subjective but complex process of inserting additional train pathes to obtain a value for capacity (Lindner, 2011). The second edition (UIC, 2013) improves on these points, while still tending to over-estimate capacity because of the decomposition of station and switch areas as well as ignoring influences of train path dependencies between station areas and adjacent line segments (Weik et al., 2020). Nevertheless, the method is widely applicable and presents a fast, simple, and efficient approach (Landex et al., 2008) to assess the capacity of railway infrastructures, if a schedule is given.

2.4. Capacity assessment given realized operations

Only limited literature exists on the topic of assessment of realized operation and only a few methods consider disturbed conditions when assessing capacity (Goverde et al., 2013). Yuan and Hansen (2004) empirically analyse the capacity occupation of the station The Hague HS in the Netherlands. They compare planned to realized capacity occupation and conclude that the realized occupation is higher. Furthermore, station utilization considerably affects train delays and their propagation (Yuan and Hansen, 2004). Goverde et al. (2013) assess the effect of different signalling systems on the capacity of a railway line between Utrecht and Den Bosch in the Netherlands. They extend the UIC norm 406 methodology with a dynamic infrastructure occupation and also compare planned and realized operations. Realized operation is generated using the line’s given schedule and a Monte Carlo simulation for generating initial delays. Capacity and delay propagation depend also on dispatching strategies. Disturbed conditions lead to less capacity occupation because fast trains must go slower and operations are more homogeneous, however, if dispatching focuses on minimizing consecutive delays, capacity occupation increases while delays stay low. They also find that the more efficient ETCS L2 cab signalling system is less sensitive to disturbances than the Dutch legacy system. However, they do not focus on the capacity occupation of the different train dynamics. Weik and Niejen (2020) study the effect of running time variability on line capacity on the central section of the Cologne commuter rail network, however not with realized operations or based on the schedule but rather using a queuing model focusing on finding the maximum frequency with which the line can be operated under variability of running and dwell times. They find that running time variability has a significant but rather weak impact on line capacity; the effect increases on longer corridors. Chu and Oetting (2013) study the impact of large disruptions. They find that for full closures, operations need to be rearranged, and this requires turnings. The impact of those extra operations requires capacity, which might not be available; hence some scheduled trains must be cancelled. However, no quantification of how much capacity is occupied by such turnings or how much capacity is freed by cancelling trains is presented (Chu and Oetting, 2013). Medessi et al. (2011) propose to extend schedule analysis to stochastic variations, to identify overlap and conflicts in train operations. This is more directly related to evaluating the probability of conflicts given a schedule and a buffer time than quantifying the minimum time required to run a set of services, assuming no stochastic variations of operations.

2.5. Methods for analysis of realized operations

Realized operations can also be analysed without looking at capacity. When railway operations are performed according to a schedule, computing the difference between planned and realized operations as well as analysing it to understand its performance and causes is relatively simple. The study of delay often takes a descriptive approach, i.e. transforming data into easy-to-understand graphics and indicators that a human analyst can further study for insights (Graffagnino, 2012). Quantitative studies use systematic analyses (for instance, aggregation, factor analysis), or unsupervised data analysis to automatically identify the different patterns of delay evolution (Cerreto et al., 2018), as well as probability fitting of specific distribution forms to the realized data (Yuan and Hansen, 2004). Delay descriptions are useful to understand the extent of phenomena, and possible correlations between system processes and system performance (Krüger et al., 2013; Minbashi et al., 2021).

2.6. Research gap

Capacity assessment with or without predetermined, scheduled operation has received much attention, but an a-posteriori recorded capacity assessment of realized operations has not yet been performed. In this paper, we define and quantitatively evaluate the steps to
undertake such an analysis, which complements a-priori capacity assessment (with planned operations, and no deviations from schedule or operational variations) and common a-posteriori analysis (with realized delays, but no study of capacity occupation of specific blocking time components or train dynamics).

A specific way to include variation in the blocking times is by including stochasticity in prescriptive formulas (Stok, 2008) fitted to reality (Medecossi et al., 2011) which has been so far used to identify the extent of possible blocking time overlaps. Like the approach proposed in this paper, Medecossi et al. (2011) can consider both recorded or simulated data and even a mix of the two. To the best knowledge of the authors, capacity assessment of simulated train trajectories with operational variation to analyse railway operations has been proposed only by Goverde et al. (2013), but only for prescriptively simulated operations, and not for realized operations. Moreover, the analysis by Goverde et al. (2013) focused on capacity occupation as a whole, i.e. not on the imputation of the contribution of train dynamics on capacity occupation, and not on the identification of latent capacity critical operations only occurring due to deviations from schedule.

Furthermore, heterogeneity in train sequence strongly impacts capacity occupation (Abril et al., 2008; Khadem Sameni and Moradi, 2022). A heterogeneous train sequence is a train sequence that comprises trains of different speeds or stopping patterns. A homogeneous sequence instead comprises trains that have roughly the same mean speed on the investigated infrastructure. The deviation of trains’ speed and thereby occupation time in a heterogeneous sequence in real operations can further impact capacity occupation. However, detailed effects have not yet been fully described and Khadem Sameni and Moradi (2022) state in a recent literature review that developing approaches for quantification of train sequence heterogeneity is necessary. The approach presented in this paper allows us to assess capacity a-posteriori making it possible to study the impacts of heterogeneity on capacity occupation in detail.

This paper’s most important contributions to the literature on capacity assessment and operations analysis are.

- We use realized data to derive a description of operations with deviations from schedule including day-to-day variations. On such a description of operation, we perform capacity assessment and study the variation in outcomes. This pertains to variation in total capacity occupation, location of capacity critical sections, identification of latent capacity critical sections and contribution of different blocking time components, and train dynamics making up the total capacity occupation.
- We quantify, on a specific test case, how capacity occupation with realized deviation from schedule differs from capacity occupation as planned in schedule. The magnitude of the distribution tails specifically describes how much capacity is lost simply due to day-to-day variations in operating conditions.
- We consider the heterogeneity of the planned schedule and show how it influences capacity occupation by expanding the bottleneck concept of UIC norm 406 to a critical path. By combining the analysis of heterogeneity and variability in operations with this expansion, we identify what insights can be generated from such an assessment and how they can be relevant for decision-making.

3. Approach

3.1. Constructive approach

This research focuses on how trains occupy capacity in planned and realized operations. Hence, approaches to assess capacity which only implicitly consider train movements as analytical formulae, queuing theory, parametric methods, and partially also optimization methods are unsuitable. Furthermore, we study planned and realized operations, thus a schedule is available and there is no need to rely on methods that are designed to assess capacity without a given schedule. Therefore, we use a timetable compression approach with data from realized operation.

This approach has limitations which have already been identified in the literature (see Landex et al., 2008; Lindner, 2011; Meirich, 2017). The main limitations are the dependency on having a schedule, the limited assessment of quality (i.e. stability and efficiency of the schedule), and neglecting delays. By assessing realized operations including delays and comparing to the planned operation (from the schedule) we overcome these shortcomings. However, note that these shortcomings remain valid if only planned operation is assessed, e.g. if operations data are not available.

We use an open data source that describes only recorded departure times from stations with a planned stop and does not provide additional details at operational points (e.g., stations where the train is only passing; important junctions, etc.). We assume the data recorded is correct, and there is no bias in time or space (see instead the approach proposed by Goverde and Hansen (2000), to counter a similar bias encountered elsewhere). There is also no information about the status of the signalling system, nor conflicts or dispatching actions, apart from what can be reconstructed from the available station data.

3.2. Reconstructing detailed trajectories from sparse observations

The lack of more detailed data means that the first step to analyse operations is to reconstruct the missing information to have a description of operations in time, space, and speed with sufficient resolution. This chapter describes the approaches proposed for this purpose. The technical implementation and the case study are described in Chapters 4 and 5.

The available data leads to two challenges: (1) reconstructing the planned operations, which is the trajectory corresponding to the schedule, assuming no deviation in departure and arrival times; and (2) reconstructing the actual trajectory (i.e. what happened in the realized operations). Compared to the detailed, uncertainty-aware reconstruction performed by Sessa et al. (2019), the much less complete set of data available for this research requires assumptions on the trajectory of trains between two measured points. For both
cases, the problem is underdetermined, in the sense that multiple microscopic trajectories would fit the macroscopic data of the recorded time points. Among this set, we choose the trajectories that can be computed according to the principles outlined below.

Importantly, we assume that we can separate the trajectory estimation problem between stops (i.e., that the speed performance and the trajectory on a specific station-to-station interval does not influence those on another interval). In planned operations, the train running times between any two stops are calibrated to have arrivals and departures matching a realistic dwell time (depending on train type, at least around 1 min; we do not impose a strict threshold), and the planned departure time (no early or late arrival). In other words, the trains use running time extensions in our simulation to possibly run a bit slower, have an arrival at the planned time, and do not always run as fast as they could or are allowed to. In cases where there is a difference between the fastest running time and the planned running time (i.e. a running time extension), the trajectory will be modified towards slower speeds, in which case the entire speed profile is proportionally slowed down by the ratio of total running time to running time extension. A similar procedure is performed using the realized departure time, for the realized trajectories.

In any case, this proposed approach cannot distinguish between delayed trajectories and conflicts, nor the actual location of a conflict in cases where a train is delayed. For instance, a train might go faster for part of the distance and then face conflicts or go slower for another part and face no conflict but arrive at the same time. While delays can be recorded in the realized operations, their cause is unknown from the available data. Possible causes include for example dwelling longer in stations because of increased demand, on track stopping because of signal failure, or running slower because of wheel slip due to heavy rain.

The results of this step are full trajectories of each train, which being synthesized, can be described in very high detail in terms of time and space. We consider a resolution of 1 m and 1 s. Given the location of the signals, blocking times can be computed along a trajectory. Please note that the intended analysis can also be performed on data with lower resolution, e.g. 5m and 5s which shortens the running time of the trajectory reconstruction. For this study, we use a trajectory reconstruction software that was designed for handling moving block operation and therefore required very high precision.

3.3. Capacity assessment approach

After developing assumed trajectories of all trains and calculated blocking times for each block section, the trajectories are compressed to assess capacity using a method similar to UIC norm 406. We adjust the method because the number of trains along a line

The capacity is thus determined by the used and lost time of block occupation on the trajectories over the whole line section. \( C_B \) is the maximum number of trains that can run on each block, divided by the length of the occupied period \( t_d \), the compressed period duration. It is measured from the beginning of blocking of the period’s first train until the same train in the next period. The compression is thereby always conducted over the whole line section \( B \) since the occupation of each block section \( b \in B \) is dependent on each other. To compute the capacity of a line \( C_B \), we identify the most restricting value, which is the minimum \( C_b \) over all blocks of \( B \). This determines the maximum number of trains that can be possibly running per hour on the line. If \( n_{\text{train}} \) is constant for \( \forall b \in B \), the maximum length of the occupied period \( t_d \) determines the capacity.

Because the timetable is compressed following the approach of the UIC norm 406, \( C_B \) represents the theoretical maximum number of trains per unit of time (i.e., capacity) under the specific operational conditions, including infrastructure layout, train sequence, and signalling system. By taking the whole line into account for calculating the minimum number of passing trains, the problem of where to divert lines into sections is sufficiently treated.

The proposed approach differs from UIC norm 406 in two key respects. First, it does not only consider the occupation time from the beginning of blocking of the first train in the sequence to the end of blocking of the last train during the investigated period but also the time just before the beginning of blocking of the same first train in the following period. This allows analysts to study the full capacity occupation of one period. Second, it does not insert additional train paths to fill the investigated time period. Instead, the full number of trains is extrapolated from the considered sequence. This approach implicitly includes additional train paths which would be
proportionally similar to those in the considered sequence. For this reason, a full period should be considered. This approach can slightly overestimate capacity because no check is made to determine if all of the trains actually fit into the investigated period. However, the approach is simpler and faster than the complex process of inserting additional trains in a partially subjective fashion (Lindner, 2011). Finally, the approach represents an achievable capacity by conservatively rounding off the resulting values.

The proposed approach is based on a periodic schedule. This means capacity does not depend on the arbitrary moment where the period is assumed to start and end, as long as the full period of time is considered (including the time between the last train of one period to the first train of the following period). We call the arbitrarily chosen moment in time where the period starts the splitting point, and the period considered is the period from a splitting point until the same offset in the periodic schedule. For each and every splitting point the compressed period duration and $t_d$ are equal. The sequence of trains does not change and the influence of the train sequence heterogeneity remains the same, just at a different point in time. The capacity is thus invariant under the choice of the splitting point in periodic operation.

The proposed approach can also be applied on non-periodic operation. In this case, $t_d$ must be considered as the time between the blocking time start of two trains which mark the beginning and end of a considered interval. Since the operation is non-periodic, the interval should be sufficiently large to consider the train sequence heterogeneity throughout an operation day. For non-periodic operation, the splitting point is capacity-relevant, thus multiple splitting points should be tested.

In any case, the maximum value of $t_d$ for the minimum of all $C_b$, the compressed period duration, is not only dependent on the operation in block section $b$ (in periodic operation also multiple block sections can show the maximum value of $t_d$) but on the operation of the whole line section under analysis. We describe this phenomenon in more detail by defining a critical path.

The critical path is a sequence of operational interactions of multiple trains, over one or more blocks wherein a change of any operation would lead to a change in the capacity occupied by the train sequence. The critical path therefore includes all constraints limiting the compression of a train sequence. It includes interactions of different kinds of operations and can follow a train on its path, go backwards or move from one train to another.

An example of this concept is shown in Fig. 1a as a time–distance path describing a compressed timetable. It shows three train types running in a period, with their blocking time represented in light grey (train P), dark grey (train Q) and yellow (train R). The train descriptions include the period numbers (e.g., trains P1 Q1 and R1 are in period 1; trains with a zero are from the previous period and trains with a 2 are from the successive period). All trains go over the same block sections, which are identified as $b_1$, $b_2$, until $b_9$.

Fig. 1. Critical Path Concept for Capacity Analysis a) Compressed periodic timetable with trajectory interactions b) Critical path in schedule c) Possible latent critical path in operation due to longer dwelling of train R1 in $b_3$. 

J. Lordeick et al.
Journal of Rail Transport Planning & Management 30 (2024) 100446
example b1 and b2 are shorter than the other block sections. The trains have different speed and stopping patterns. The critical path of this operation is shown in Fig. 1b.

Four operational interactions between trains can be distinguished (operational interactions describe the evolution of train sequence heterogeneity between two trains microscopically). In the example the entire period cannot be compressed further because of the following train interactions. Train Q1 on b3 runs immediately after Train P1 on b4; and train Q1 on b9 must happen immediately after train R1 on b9 and train P2 on b9. On those occasions, the blocking times of the involved trains touch and the trains follow each other with the minimum headway. We call these compression limiting interactions.

We define a blocking time gap between two trains as the time interval between the end of blocking of the first train and the start of blocking of the following train. The blocking time gap is capacity-relevant and is linked to the headway time, but must not be equal to the headway time. During this blocking time gap, the block section is neither physically occupied nor blocked because of the signalling interlocking logic. However, it is not possible to use the blocking time gap for a faster following train without the possibility for overtaking. The literature refers to a blocking time gap in this case as lost time (Abril et al., 2008; Jensen, 2015). This lost time is capacity relevant because it determines the value of \( t_{bh} \) and thereby also the compressed period duration, it is displayed in light blue in Fig. 1b and c.

Blocks where the blocking time gap changes are especially relevant for improving capacity. We call these either separating interactions if the blocking time gap between two trains increases (for instance, b9 between train P1 and Q1) and compensating interactions if the blocking time gap decreases (for instance, b5 between train P1 and train Q1).

If the blocking time gap does not change but trains are also not running at minimum headway behind each other we call these neutral interactions. They do not contribute to the blocking time gap in the schedule or realized operation under investigation, but if the trajectory of one train changes they directly change to separating or compensating interactions and become relevant for the blocking time gap. Those are for instance the constant blocking time gap in Fig. 1a between Train P1 and Q1 across block sections b6 and b7.

The type of operational interaction can be used to identify methods for infrastructure managers and railway operators to improve capacity. In compression limiting interactions, trains are running at minimum headway behind each other and hence technical improvements on the infrastructure or on the rolling stock are necessary to increase capacity. In separating and compensating interactions, operational decisions (e.g. an express train passing a station and a stopping train dwelling) or varying vehicle characteristics can be changed to increase capacity occupation.

The critical path enriches the understanding of the effect of the train sequence heterogeneity on capacity. In general, heterogeneous traffic results in the critical path extending over at least two blocks and at least two trains. Because the example in Fig. 1 shows periodic operation with a constant number of trains the maximum value of \( t_{bh} \) is equal to the compressed period duration and the critical path is the same for every period in schedule (see Fig. 1b. The critical path can thus also be understood as defining the minimum (thus compressed) period duration for the sequence of trains in the period. We show in Fig. 1b the critical path in case of scheduled durations of train operations. In case durations would change, the critical path can change. One possible case is reported Fig. 1c (discussed later).

The train dynamics on the critical path are then the root causes which limit capacity. Using this concept, the effects of train dynamics on capacity can be computed in detail on the critical path. Furthermore, it enables analysts to identify the contributions to capacity utilisation from the schedule structure, from the train dynamics, and the length of dwell processes. The behaviour of trains along the entire critical path can be studied in detail by quantifying absolute values and shares of capacity occupation by different elements of the blocking time as well as of the train dynamics.

Oetting and Griese (2016) propose a comparable approach tailored for calculating the minimum headway between two trains. They calculate the difference in running time of both trains from every block section of the line section to the second last block section and add the headway-relevant blocking times in the last block section. The maximum time resulting from these calculations is the minimum headway. With this approach, the minimum headway of a faster train following a slower train can be calculated so that the faster train does not need to stop behind the slower train. In the case of a slower train following a faster train, the minimum headway will be only the headway-relevant blocking times in the first block section. These minimum headways can be used to calculate the capacity in a similar way as presented in the formula above if considering the minimum headway time of the last train of the current sequence to the first train of the next sequence. This is the basic principle of analytical models (Khadem-Sameni and Moradi 2022). This critical path concept represents an adaption of methods to calculate minimum headways (as for example in Oetting and Griese 2016). It takes a reverse approach to assess capacity occupation with an analysis to identify, in simple terms, why the analysed train sequence cannot be compressed any further. This procedure is similar to searching for minimum headways between trains but is tailored for a constructive approach for analysing capacity. We consider it therefore an extension of capacity assessment using timetable compression laid out in UIC norm 406 (see for example Landex 2008). It may serve as an analysis tool for better understanding capacity-relevant operations. This is described in the following section.

### 3.4. Computation of factors influencing capacity occupation

Given a detailed trajectory, the minimum headway in each block between two trains in a compressed timetable can be computed, and its influencing factors can be studied. Traffic is heterogeneous and subject to a multitude of concurrent effects, which can be identified separately.

Fig. 2 describes the components considered in the blocking time. We distinguish static components, which do not depend on the speed and the specific operations of a particular day, and dynamic components which do. Train dynamics comprise the running time as part of the blocking time (and thereby influence approach and clearing time). The blocking time components and train dynamics phases (in what follows we consider both as factors) are illustrated in Fig. 2 with different colours. They consist of.
• Blocking time components
  o Path formation, sight time, path release (depending on technical characteristics of the infrastructure and regulation, considered static)
  o Approach time (depending on signal distances and speed, dynamic)
  o Running time (depending on vehicle characteristics and speed, dynamic)
  o Clearing time (depending on vehicle characteristics and speed, dynamic)

• Train dynamics phases
  o Acceleration time (compared to a reference trajectory at constant speed)
  o Braking time (compared to a reference trajectory at constant speed)
  o Dwelling time (compared to running)
  o Cruising time (compared to acceleration and braking, i.e. running at non-constant speed or dwelling)

Fig. 3 illustrates the blocking time gap between the trajectories of two trains, which have different speeds and stopping patterns. The dynamic performance of the two trains is identical. The first train is running earlier (top of the figure) with constant speed and without stop. The second train is running later and stops once, including deceleration and acceleration. Since those trajectories are already compressed, the blocking time gap between the two trains varies over space, due to various factors. As the blocking time gap in a compressed timetable is crucial to assess capacity, we can associate a capacity effect to each train dynamic in connection with the critical path concept.

Lost time is determined by the blocking time gap between two trains caused by the sequence of operation on the critical path (see Fig. 1). The bottom part of Fig. 3 illustrates the change in the blocking time gap between two trains. The colours again refer to the blocking time components and train dynamics respectively which cause the interval change. The running time is composed entirely of the train dynamics and hence the left part of Fig. 3 is a further detailing of the right part.

For example, in the middle block section, the second train’s planned stop increases the blocking time gap (the time difference is positive and depends on braking and dwelling). In the following block section, the second train accelerates and therefore is still slower than the first train. Both operations are separating interactions. Afterward, the trains run again at the same speed and the blocking time gap does not change (neutral interactions), however, the already generated blocking time gap remains and occupies capacity as lost time.

In general, any dynamic component of blocking and all train dynamics can vary for every operation. Empirically, the effect of varying operations in real-life is mostly attributable to changes in the running time component of blocking time. Overall, this results in a probability distribution of blocking times for each train in every block section. We can bound those probability distributions by a minimum and a maximum so the analysis excludes extreme outliers that need to be assessed differently because of their potentially extreme cause not explainable by regular day-to-day variations.

In short, the trajectory of trains varies in real-life operations, which means that the critical path and its influencing factors may change, and their grand total, namely the compressed occupation time \( t_{db} \), and the minimum value of \( C_b \) can also change. In other words, the critical path, the capacity occupation in each block, and the capacity of the whole line section can differ between actual realization and planned operation due to the differences in the trajectories. When assessing capacity using the critical path the impact of multiple trajectories compared to the planned one can be characterized by two prototypical cases, as follows.

First, it is possible that the critical path is topologically always the same but the operational interactions vary to some extent, corresponding to the accumulation of the probability distribution of all trains. In other words, over a line, the entire set of trajectories shifts stochastically, but \( t_{db} \) is always constrained over the same critical path. The distribution of capacity is thus a stochastic univariate value, which replaces the deterministic value calculated in the UIC method. The potential railway insight is still to address the conditions that make the critical path appear. The critical path would be similar to Fig. 1b, i.e. same sequence of trains over the same block sections.
Second, the varying blocking times may cause different latent critical operations and critical paths at different probabilities. Specifically, multiple paths and different operations can be critical for different actual operations. Normally the path which is critical under scheduled operation remains one of the critical paths under realized operations. The other possible critical paths can be similar; for example, compression limiting interactions occur in block sections preceding or succeeding the block identified using the schedule. This is often the case if speed differences are the determinants of the critical path. Otherwise, the other possible critical paths can be at completely different locations; for example, if they are caused by a different stopping pattern or the branching or merging of a line. The potential railway insight in this case is that putting resources into fixing specific operational interactions might not improve the situation much since other problems need to be fixed as well. Assume that operation of train R1 on block section $b_3$ is longer due to slower deceleration, longer dwell time, etc. If this is sufficiently long, the blocking time gap between train Q1 and R1 over block section B3 would reduce to 0 and a new blocking time gap will open in block section $b_9$. The critical path will change to the one reported in Fig. 1c and the operational interactions from $b_3$ to $b_9$ between trains Q1 and R1 become part of the critical path.

4. Experimental setup

4.1. Overall approach

The proposed method reconstructs trajectories and determines blocking times using an event-based microsimulation. The movement of every train on 1 m of track is considered an event; events are linked by the standard equations of motion, given the maximum allowed acceleration, cruising speed, and braking of vehicles. For each second the type of train dynamic of the train is recorded (i.e., cruising at constant speed, braking, accelerating, or dwelling). The simulation is asynchronous and parallel, with the additional constraint of matching the recorded macroscopic times from the schedule or the available data on delays at stations. Fig. 4 illustrates the approach. The algorithm basically follows a given path in the network (blue boxes) but makes train specific calculations at every metre (red boxes). Generally, the structure of the asynchronous simulation limits the possibility of dynamically influencing trains while running. The path, stops and dwelling times must be specified beforehand. On the other hand, it can be easily parallelized. The blocking data can be checked at once so that the algorithm ensures that blocking times do not overlap each other.

The result of the simulation is a table that contains the start time of formation, sight, and approach, the entrance of the head of the train which marks the start of the running time, the head exit which marks the start of the clearing time, the tail exit which marks the start of the release time as well as the release end time for every block. Additionally, the braking, accelerating, dwelling, cruising and cruising with approach speed while approaching a signal (for line signalling in operation on the case study corridor) times are stored for each block. With these data, the capacity occupation of every block and different train dynamics can be studied in detail. The analysis and the microscopic reconstruction of the trajectories are separated modules, so having detailed trajectories from another source (e.g., Medeossi et al., 2011) would allow the same analyses.

4.2. Description of realized operations data

In order to obtain operational data based on real-life, sequences operated in the past are re-simulated aiming to generate trajectory data (as outlined in Chapter 3). In this process the scheduled departures are adjusted by extracting and inserting the realized departures gathered from the Swiss open transport data platform (SBB Business office SKI, 2021). The datasets contain every scheduled departure and arrival of each passenger train line in Switzerland, at the resolution of minutes. The actual departure is reported measured in seconds, actual arrivals are not used for operated trains. Days with cancellations and deviating paths are excluded because they would bias the result. Freight trains are not recorded in the scheduled data, nor in the realized data; their operations are therefore not considered. A future extension, provided that data are available, should address this aspect. For the case study, four months (January to April 2021) were used which results in 118 days with the same 2 h of operation were simulated. Two days were excluded because of incomplete data. Since the approach is based on actual operations data it is closer to real-life operation than for example a
Monte Carlo simulation generating initial delays as used by Goverde et al. (2013).

4.3. Implementation

The approach focuses on one direction of travel if train paths do not interfere with each other (e.g., on double track lines with one directional operation). In the case study the trains in direction Chur-Zürich were considered only for what pertains the small single-track section in a tunnel between Sargans and Ziegelbrücke, where they interfere with the opposite direction. Those counter-flow train
trajectories are bound to the part close to the station closest in front and behind the single-track tunnel. Other single-track sections in stations are not crucial or are too simplified for capturing robust effects. The case study line section operates with line signalling. The signal distance was measured as exactly as possible from Swiss aerial pictures. The main signals of previous block sections serve as advance signals instead of additional advance signals. The train simulation and recording of approach times are adjusted accordingly. The behaviour of signals follows the normal applications of blocking time theory and interlocking.

The approach speed for red signals is set to 40 km/h which allows safe braking also for freight trains with poor braking performance. However, on the S-Bahn corridor, the application of this approach speed renders the operated S-Bahn sequence impossible. Due to the missing information on the case study, it is unknown how the quickness of operation is achieved. Hence, for the simulation, the approach speed for S-Bahn trains is set to 80 km/h which is still realistic because of better braking characteristics. This change allows to keep the schedule. Under practical conditions, the stopping signal at a station normally shows green if the block ahead is unoccupied and hence there is no forced braking of the S-Bahn which reduces the approach time dramatically.

As described in Section 3, the train trajectories for the actual operation are modified to match the recorded data on departure delays at stations (arrival delays are unknown). Therefore, the trains are proportionally slowed down or speeded up if they built up/or reduced respectively a delay between two stations. These delays can be caused by conflicts on the infrastructure, requiring a train to run slower. Since the arrival of the train at a station is unknown, we impute it by considering multiple causes, more specifically trains dwelling longer than planned at stations, for example, due to high demand or conflicts with other trains. We adjust the modification of running time stochastically so that the delay is caused sometimes because of longer dwelling times and sometimes by being slower or faster than planned on track sections between stations. This procedure might not always correctly re-simulate the cause for delays, but it guarantees that several potential reasons for delays are produced in the data.

The resulting train trajectories represent a simulation result that is as close as possible to the truly realized trajectories. Because of simplifications and the stochastic behaviour, these sequences might be partially invalid because of incorrectly computed blocking times in intermediate sections along the path. Nevertheless, shifting train trajectories to make them fit into the sequence, or imputing different trajectories without any additional information would reduce the simulation accuracy compared to actual operations and therefore bias the analysis results more strongly than keeping overlapping blocking times. For capacity assessment, the sequence is rearranged by compression therefore overlapping trajectories have no influence on results. The blocking times are calculated first for each train separately and are merged in a later step. Since buffer times are not considered for capacity occupation, overlapping trajectories have no influence on the analysis after compression. The pure process of compressing the trajectory data is conducted as suggested in UIC norm 406 (UIC, 2013). Thereby only trains are included which have their first appearance inside the investigated line, inside the studied period to avoid double counting. The analysis is then performed according to the approaches described in Chapter 3. Simulation and analysis are performed on a standard laptop machine (i7, 16 GB Ram).

5. Case study

The mainline railway line between Zurich and Chur serves as the case study of this research. It is a typical mainline railway with mixed passenger traffic of different speed and stopping patterns as well as multiple international trains. The line between Zurich and Chur has a length of roughly 115 km and includes six line sections and four stations that have relevant diverging lines and changes in operation (not including Zurich HB and Chur). Overall 31 stations with and without junctions are served. The line includes different infrastructures from long tunnel sections with a maximum speed of 160 km/h over S-Bahn sections with very little station distances to a single-track tunnel. Therefore, it serves as a diverse example for capacity assessment in real-life railway operations. Moreover, for this case study, open data on delays at stations are available.

Fig. 5 shows a simplified topology for the corridor. From Zurich HB the corridor is separated into a branch for local trains running along Lake Zurich and the Zimmerberg Base Tunnel for regional and long-distance trains. Both lines merge ahead of the station Thalwil. South of Thalwil the line follows Lake Zurich and continues to the Walen Lake passing major diverging stations Pfäffikon (for simplicity, we refer to the Station Pfäffikon SZ; also when SZ is not explicitly mentioned) and Ziegelbrücke. The line merges in Sargans with a line from St. Gallen and Austria and continues to Chur. For the analysis, the line is separated into the following six line sections and four station areas: Lakeside S-Bahn route, Zimmerberg base tunnel, station area Thalwil, Thalwil-Pfäffikon, Pfäffikon station area, Pfäffikon-Ziegelbrücke, Ziegelbrücke-Sargans, Sargans station area, Sargans-Chur. Zurich HB is excluded from the analysis because of its very simplified modelling; the analysis starts after the station area.

To build the simulation model, fully based on open data, we used public cartography of the area; aerial photography detailed enough to allow localization of ETCS balises, and videos of cab rides (see Cab Ride Zürich to Chur, 2015; Führerstandsmitfahrt in wundervoller, 2019). Stations are simplified to include only the necessary or easy-to-identify used tracks. Stations and junctions are only modelled to the necessary for train movements in stations. The maximum allowed speed is taken from Open Railway Map (see Open Railway Map, 2021). Slope, curve radius, and other parameters are neglected.

Schedules were obtained from Fahrplankelder.ch (BAV, 2021), and adjusted for replications throughout a regular interval schedule. Furthermore, the rolling stock is assumed to be the typical one used for those services, for which public sets of parameters are available. Train dynamics are sensitive to these parameters, especially engine parameters determining braking and acceleration capabilities. However, since the case study is bounded by the schedule and the realized operation data, we analyse this sensitivity analytically only.

The operating pattern in Switzerland is periodic with a period length of 1 h, with only minor changes between periods. For this case study analysis, the period between 3 p.m. and 4 p.m. was modelled. This period faces high demand but does not include extra S-Bahn trains on the Zurich network. We also consider the irregularly operated Railjet Express (RJX) service coming from and leaving for Austria. To capture all trains realistically the simulation starts at 2 p.m. with 1 h of “warm-up”.
6. Results

The following chapter presents the results of the case study analysis with the approaches outlined in Chapter 3. We follow an approach-based structure to demonstrate the developed methods and report the results of the analysis of scheduled and realized operation after each other or in direct comparison per analysis approach as far as possible.

6.1. Blocking times in schedule and realized operations

The analysis starts by reporting the planned operations, as simulated by the approach. Fig. 6 shows the planned schedule, with its blocking times uncompressed. It shows the variety of traffic, with multiple trains entering and leaving the line, and the variety of stopping patterns. The grey boxes report the dynamic blocking time supplements (approach time, and clearing time, similar to Fig. 2). The remaining static supplements (path formation time, reaction time, path release time) are not displayed but are considered. This helps distinguish between trains in the figure. The coloured boxes report the train dynamics comprising the running time, differentiating between cruising at the maximum allowed speed, braking, accelerating, and cruising at approach speed. The different train categories are identified in both Figs. 6 and 7 by different colours based on the graphically underlying speed profile. In Fig. 7, because

![Fig. 6. Blocking time diagram of planned operations.](image)
of the grey colour scheme, the train types can be better distinguished. S-Bahn and regional express (RE) lines have a red speed profile and EC, RJX, ICE, IC and IR lines have a blue speed profile because those lines are simulated with the same rolling stock characteristics. The train types can also be distinguished by the first digits of the identifier (s and re or ec, rjx, ice, ic and ir respectively).

Fig. 7 reports the distribution of realized blocking times. The medium grey boxes report the range from the 10th percentile to the 90th percentile, while the median duration of each blocking time is shown in black. To allow a direct comparison with Fig. 6, the light grey reports the scheduled blocking time. Overall, most blocks have a realized blocking time that is larger than scheduled; and for some of them, the black (indicator of variation) is particularly prominent. Also note that some services (especially those without stops) run on the planned trajectory with high probability. The area between Thalwil and Pfäffikon, around minutes 0 to 20 shows a large dark grey stripe, corresponding to highly irregular train operations by an S-Bahn service.

6.2. Spatial analysis of capacity occupation in schedule and realized operation

The next step is to compare the capacity occupation for planned and realized operations as shown in Fig. 8. Each line section is considered separately, as described in Chapters 3 and 4. On each line section, the traffic volume changes significantly (due to the ending of services, new services, or merging and diverging of services). The vertical axis reports maximal capacity occupation as a % share of the compressed operations versus the total time available in the period (60min). The red lines show the planned operation. The realized operations are shown as box plots in black: box boundaries define the 25th and 75th percentiles; the heavy black line displays the median; the whiskers the largest value no further away than 1.5 times the 25–75 percentage range, while the outliers are reported as single dots. The capacity occupation is a static value for the schedule (reported in red horizontal lines in Fig. 8) but forms a distribution for realized operations. The scheduled capacity occupation has its largest value at 36%, for the station area of Sargans. However, major stations with diverging line sections are only considered in a simplified manner, hence their capacity is likely underestimated. The largest non-station capacity occupation of roughly 30% is the lakeside route between Zurich and Thalwil which is mainly used by S-Bahn services.

As shown in Fig. 8, a high share of the mean and distribution for capacity occupation in realized operations lie well above the scheduled value. For instance, the line section between Sargans and Chur has most data points above the scheduled value; and the median in this line section lies about 3 percentage points higher. This is an indication of variability between operations, suggesting that realized operation generally occupies more capacity than scheduled operation. The mean value of this capacity occupation in realized operation as well as the outliers which often far exceed the 75th percentile of the distribution are both still below the maximum proposed capacity consumption values from the UIC norm 406. This means that from a capacity perspective, more trains could be operated on the investigated line. This available capacity might be occupied by freight trains in real-life, which we do not model here. Furthermore, schedule restrictions because of planned connections in major stations further limit the usability of available capacity on this line. Looking at the pattern of the distribution along the x-axis, both scheduled and realized operation follow similar patterns.

In short, consistent with a fundamental theorem of scheduling (Mohring, 2001), the capacity occupation in realized operation is

![Fig. 7. Blocking time diagram based on the reconstructed trajectories.](image-url)
always larger than the deterministic occupation in planned operation. For a set of scheduled events, interconnected as a series or parallel, each event has a probability distribution of completion time. The theorem states that the average completion time of the sequence of events is larger than the completion time of the deterministic series of events of the average duration. This helps explain the underestimation of expected event duration frequently observed in practice when durations are subject to operational variations. In our analysis, we refer directly to the planned operation, which does not need to be equivalent to the average duration of events. In fact, buffer times are used differently, to increase chances of on-time arrival at specific stations, so the picture is even more complex here. This opens the possibility to study the complexity of the chain of critical operations, how many activities occur in parallel or in series, and how heterogeneity affects them. We do not perform such an analysis in detail but do analyse how heterogeneity plays an

Fig. 8. Planned (red line) and realized (box plots) capacity occupation over each segment of the corridor.

Fig. 9. Compressed blocking time plots: identification of the heterogeneity contribution of each train (left) and the chain of critical operations (right).
important role in operations for the successive sections.

The following sections focus on the Sargans-Chur line segment because its infrastructure and train sequence are simple enough to easily display detailed train behaviour, while the heterogeneity in its operations is big enough to serve as a test case for our analysis.

6.3. Train sequence heterogeneity and capacity occupation in scheduled operation

Fig. 9 (left) shows a compressed blocking time plot of the planned operation, with the occupation times highlighted in orange. Each train contributes to capacity occupation by lost time because of train sequence heterogeneity. The blocking time gaps are reported in grey and black. The colours are different for distinguishing between two trains, since we assume the blocking time gap between two trains is attributed in equal amounts to the train before and after because both contribute to heterogeneity in the sequence. Both trains fulfill a specific purpose in the schedule, none is more responsible for heterogeneity than the other. In general, much of the lost time is the result of different stopping patterns and speed profiles of the trains in the sequence. This analysis attributes lost time to specific trains and thereby allows to identify optimization potential for changing single train services and their sequence.

Fig. 9 (right) shows the critical path of the planned operation. The same vertical and horizontal axes are used as in the left part. All occupation times are now displayed in light blue, and the critical path in dark blue runs along the operations. The capacity is determined by the operational interactions along the critical path. Those which contribute strongly to lost time are between the ICE and the S12 (the first two trains with two similar trains following at the end of the sequence). There are three separating interactions: first at track metre 92068 (entrance towards the Bad Ragaz station), second at the track metre (Bad Ragaz station) and third the stop of the S12 at track metre 96468 (in Maienfeld). The line also has one compensating interaction at track metre 100733 (Landquart station) where both trains stop with the ICE dwelling longer than the S-Bahn. In this blocking section the two periods would touch each other. All the other trains in the middle of the sequence only have compression limiting interactions in the station of Landquart along the critical path. Operations not included on the critical path are thus irrelevant for the capacity of the line section in planned operation. These operations can be changed (to a certain extent) without influencing the capacity of the line section.

The lost time is the result of the train’s behaviour throughout the entire critical path. The lost time between the ICE and S12 is the result of the speed difference in several block sections and the additional S-Bahn stops in Bad Ragaz and Maienfeld. Stopping also requires deceleration and re-acceleration which result in further heterogeneity and occupation of capacity. Hence, to reduce capacity occupation, the operation of S12 or ICE along the critical path must be adjusted. However, note that adjusting speed profiles and stopping patterns may change the critical path and therefore involve other trains. This adjustment process can be repeated either until all blocking time gaps on the critical path vanish and all interactions become compression limiting interactions (from thereon, capacity can only be increased with technical improvements to the infrastructure or the safety system) or until the critical path itself changes.

6.4. Evolution of the blocking time gap in realized operations

The level of detail available in microscopic blocking time data makes it possible to expand this analysis to determine which components of blocking time and which train dynamics contribute to the difference between scheduled and realized operations in capacity occupation. Fig. 10 zooms into the train-to-train sequence between the ICE and S12 to enable more detailed analysis of the blocking time gap between them (the figure is comparable in axes and principles to Fig. 3).

Fig. 10 presents the blocking time gap between those trains in seconds (y-axis) for each block of the section identified by the starting meter of the block (x-axis). The figure has the same x-axis as Fig. 9, focusing on the critical path line section (starting before Bad Ragaz to station Landquart). Fig. 10 shows the scheduled blocking time as a red line; and the distribution from realized operation as boxplots. The combined display of scheduled and realized operation allows analysts to see how the blocking time gap deviates from schedule in realized operation and how this might cause the critical path to change.

The operation in the third block section is the compression limiting interaction and forms the start of interactions which determine the critical path. Since the blocking time gap in the next block section is increasing it is also a separating interaction. The operations in the two block sections to the left are not on the critical path, they can be changed to a certain extent without influencing capacity. In the two block sections to the right scheduled and realized operations coincide, which shows that train trajectories are stable in these block sections. For all other line sections in Fig. 10 the blocking time gap is strikingly lower for realized than for scheduled operations. In other words, in realized operation, both trains are closer together in a compressed timetable in those block sections than the planned operation would suggest. The difference between the median of the realized operation and the scheduled operation is around 20 s, and all third quartiles are still lower than the planned case. The major separating interactions in Bad Ragaz and Maienfeld are visible as well as the compensating interaction in Landquart which together mostly determine the lost time. This diagram shows how blocking time gaps and consequently, lost time between successive trains, changes in realized operation. It thereby shows how variability and heterogeneity of operations are the main causes of capacity differences between planned and realized operations.

Fig. 10 shows that there is little possibility for operations in the third to fifth block section to become strongly separating in realized operation. The blocking time gap in the compressed timetable is small in both the scheduled and realized cases. For capacity assessment with timetable compression, it does not matter if the S-Bahn is running late behind the ICE, as long as it runs at maximum speed. Hence, even if an operation in a preceding block causes delay and a small increase in headway, the blocking time gap will stay low. Only after the first additional S-Bahn stop in Bad Ragaz, will an increase in delay by the S-Bahn become capacity-relevant. The middle and right sections of the diagram show this effect. The variability of blocking time gaps is greater here. When looking at the major separating interactions, both have little variability. However, the majority of the distribution lies below the scheduled blocking time gap, and thereby the blocking time gap in realized operation is often lower than in the schedule. Given the insight from Fig. 7 that
the S-Bahn is regularly late, we infer that the S-Bahn real-life operation fits, in most cases, the operation of the ICE better than in schedule. This can be for example the result of driving the S-Bahn at an increased speed if it is running late.

We expect the variability not to be large enough to change the critical path for these two trains significantly. First because of the relatively low variability on the y-axis and second because of the seemingly low variability on the x-axis and hence a relatively stable headway between both trains for each operation case.

6.5. Analysis of blocking time components and train dynamics for understanding changes in the blocking time gap

A final analysis on the scheduled operation considers how the blocking time gap between the two example trains can be imputed to a variety of factors, and how the heterogeneity, in planned operations, plays a strong role. Fig. 11 shows the contribution of each factor to the change in blocking time gap (the figure compares each operation in the block section to the preceding) of the block sections compared to the deterministic, scheduled case. Again, the reference theoretical background is graphically presented in Fig. 3.

Fig. 11 shows together with Figs. 9 and 10 why the S-Bahn loses (and in some blocks makes up) time against the ICE. This allows planners to identify factors which can be changed to increase capacity. The horizontal axes are the same as in the previous Figs. 9 and 10.

The left side of Fig. 11 presents train dynamics (dwelling, acceleration, and braking), and the right side the blocking times (including running time). Therefore, the left side can be seen as a detail of the right side. Fig. 9 shows that the critical path spans from the third block to Landquart. A change in the blocking time gap is visible in most block sections, however, the two separating interactions in Bad Ragaz and Maienfeld, where the dwelling time (of the S12) increases the blocking time gap, are particularly prominent. The operational interaction in the last block section has a negative contribution to dwelling time because the S-Bahn has a shorter dwelling time than the ICE and brakes faster. This difference causes a compensating interaction. We compare the running time elements between both trains, hence the braking time difference between S-Bahn and ICE at S-Bahn stops is large as well because the ICE is running through. The same applies to the acceleration time in the following block. Thus, the times reported in this diagram can be greater than the running time change reported in the right part of the figure.

In the right part, it is evident that except for the blocks immediately adjacent to the stops, there are no large contributions to the blocking time gap variations from running, approaching, or clearing time differences. The blocks in between have instead small positive and negative variations which are caused by differences in train length, running time margin, as well as different braking and acceleration performances. The effects at least partially counteract each other.

6.6. Sensitivity of capacity on train dynamics

The case study results show that the capacity occupation is sensitive to the train dynamics on the critical path because train dynamics determine the blocking time gap change. If the maximum speed is limited by the infrastructure, then adjusting braking and acceleration performance as well as the dwell time are crucial to improve capacity. In the example above, stops of one train interacting with the passing of the preceding train were especially important in causing relevant separating interactions. Dwell time reduction is however less linked to operational decisions or vehicle performance assuming that it is already set at the minimum to allow robust

---

3 This analysis could also be conducted on realized operation focusing on the distribution of blocking time components and train dynamics in single operations or interactions. Because the realized operation is fitted to reality by a running time margin imputed on the blocking time and not by adjusting the train dynamics, the case study data is not suitable for such an analysis.
passenger deboarding and boarding. Reductions of dwell time are therefore related to improvements in passenger flow during the deboarding and boarding process. Harris et al. (2022) provide a current study of influencing factors and thereby show which interventions reduce dwell time.

Nevertheless, every stopping activity involves acceleration and braking which are directly related to vehicle performance and can therefore be altered by changes in rolling stock. Although the capacity occupation is sensitive to these two train dynamics, performing a full-scale sensitivity analysis for the case study does not provide realistic insights because the case study simulations are fitted to the schedule and the realized departure times respectively. Changes in rolling stock would show either no changes in capacity occupation if we keep the fitting or would show biased results if we overrule the fitting because we deviate from reality.

Rolling stock influence on capacity must have been already investigated during the timetable planning phase, therefore we present an analytical model to show the interactions of rolling stock performance on capacity by modelling the increase of the blocking time gap between two consecutive trains in a station if one train stops and the other train is passing. The case study shows that the change of the blocking time gap because of different stopping patterns is the most relevant regarding capacity. Equation (2) introduces the relationship of the braking, dwelling and acceleration train dynamics with the blocking time gap while Equations (3)–(5) describe the underlying equations of motion. The formulation is adopted from Coviello (2015) and enriched by equations from Wende (2003).

\[ \Delta n_{p,s} = \Delta n_{1,p,s} + \delta_{Bn} + D_s + \delta_{An} \]  
(2)

\[ \delta_B = \delta_A = t_{cs} = \frac{d_C}{v_P} \]  
(3)

\[ d_C = \frac{v_s^2 - v_t^2}{2a_s} \]  
(4)

\[ t_{cs} = \frac{v_t - v_s}{a_s} \]  
(5)

\[ \Delta n_{p,s} \] Blocking time gap after station \( n \) of the passing train \( p \) and the stopping train \( s \).

\[ \delta_{Bn} \] Blocking time gap increase caused by braking at station \( n \).

\[ D_s \] Dwelling time of the stopping train \( s \).

\[ \delta_{An} \] Blocking time gap increase caused by accelerating at station \( n \).

\[ t_{cs} \] Time for the change in speed of train \( s \) (Braking or Acceleration)

\[ d_C \] Distance for the change in speed of train \( s \) (Braking or Acceleration)

\[ v_P \] Velocity of the passing train \( p \).

\[ v_s \] Current speed of train \( s \).

\[ v_t \] Target speed of train \( s \).

\[ a_s \] Acceleration of train \( s \) (Positive for acceleration, negative for braking)

**Fig. 11.** Imputation of reasons for blocking time gap change between compressed train paths.

**Fig. 12** shows a plot of blocking time gap change because of speed change (\( \delta \)) under varying speeds of the passing train (which is also the speed the stopping train must reach again) and varying acceleration or braking performance. The block section lengths are
considered constant. While braking can be assumed to be linear in reality acceleration of real trains is only linear until a specific speed at which point the acceleration decreases forming an acceleration hyperbola. However, such an acceleration profile can be transformed into a mean acceleration profile and inserted into Equation (4). The simulation described in Chapter 4.3 is implemented with non-linear acceleration (see Wende, 2003). We explore a wide range of acceleration and braking rates, which go beyond the current values of running vehicles on the specific line.

The plot shows that with increasing speed, low braking or acceleration performance can cause increases in the blocking time gap. A service brake is usually done with $0.7–0.9 \text{ m/s}^2$ (Powell and Palacín, 2015). For braking alone, the blocking time gap then increases roughly between 50 and 100s for speeds of the passing train between 80 and 160 km/h. In this scenario a dwell of 60s causes the blocking time gap to increase between roughly three and 5 min also including the acceleration time with the same performance bandwidth.

Imagine the example of the ICE and the S12 from above in a planning stage (see also Fig. 12). The S-Bahn has an additional stop in Bad Ragaz, the passing speed of the ICE through Bad Ragaz is 130 km/h. The S-Bahn has a braking capacity for a service-brake of $0.9 \text{ m/s}^2$. Braking alone results in the blocking time gap to increase by 72s. If we assume a 10% stronger or weaker braking performance, the increase reduces to 65.7s (-9%) or increases to 80.2s (+11%). If the passing speed of the ICE reduces or increases by 10%, the blocking time gap change reduces or increases linearly by 7.2s (+/-10%).

This example illustrates that with higher speeds the increase in blocking time gap generally increases while the cross relation between speed and braking or acceleration performance also increases. Especially for mixed traffic lines with high speeds of passing trains, stopping services cause large blocking time gaps which can be reduced by increased braking or acceleration performance. Powell and Palacín (2015) find that quasi static accelerations of 1.4 m/s$^2$ are still acceptable for rail transit passengers while especially the jerk (the change of acceleration or braking) influences the passenger comfort during acceleration and braking phases. Especially with advances in ATO with optimised and faster braking and acceleration profiles minimizing the jerk, this does not only improve passenger comfort, energy consumption and journey times (Powell and Palacín, 2015) but also capacity on lines with heterogeneous traffic.

This perspective makes it possible to derive simple rules and estimations for planning, however, it already shows the analytical approach’s limitations. The interaction of multiple trajectories over a line section with multiple stops and multiple block sections of varying lengths and in the varying conditions of actual operations results in a very complex analytical model. Instead of proposing many theoretical assumptions that need to be checked for a full analytical a priori-study, we resort to have this complexity exposed by the realized operations in this paper.

Nevertheless, incorporating a full-scale simulation-based sensitivity analysis into the schedule planning stage combined with the critical path approach described above can identify operations where investing in rolling stock with increased braking and acceleration performance can improve capacity. Similarly, lines can be identified where less powerful and therefore cheaper rolling stock can be used. A comprehensive sensitivity analysis would also allow studying the impact of changes of maximum speed in relation to the train sequence and the structure of blocking sections to identify infrastructure elements which would be most beneficial to optimize for higher speeds (e.g. curve straightening or use of high-speed switches).

While the analytical tools presented in this paper would be able to perform this analysis, we do not think it is a relevant addition to the paper. The degrees of freedom would be too many; the scientific value limited; the space and figures required to report such multi-level sensitivity, large; and the match to the core contributions of the paper (i.e. the estimation of actual usage of capacity) too weak with regards on the space required to explain it.

![Fig. 12. Sensitivity of the blocking time gap change to varying line speeds and acceleration or braking performance.](image-url)
7. Conclusions and future research

This work investigates the possibility of analysing the capacity occupation on a railway network including variations of realized operations. This paper presents a method based on the UIC norm 406 for calculating capacity from compressed blocking time paths, fed with train trajectories reconstructed by a simulation fitted to realized data. Next, it describes several analysis techniques for use on the compressed operations, introducing a critical path approach for capacity assessment. In addition, it describes which components of blocking time (path formation, sight, approach, running, clearing, and release time) and train dynamics (accelerating, cruising, braking, and dwelling), are responsible for capacity occupation, and how they can be influenced to reduce overall capacity occupation. While the case study data are unsuitable for a full-scale sensitivity analysis of different vehicle performance, we present an analytical model of the blocking time gap’s sensitivity to the speed of the passing train and the vehicle characteristics of the stopping train in a station. This analysis shows that improving acceleration and braking performance can reduce blocking time gap increases in train sequences with mixed stopping patterns and thereby reduce capacity occupation.

The research shows that the methodology based on reconstructed trajectories, or precisely measured trajectories if available (for instance in Medeossi et al., 2011) enables additional insights. As opposed to schedule planning this a-posteriori capacity analysis (after recording of operations) allows including the impacts of real-life operations helping to understand the impact of variability. The case study analysis shows that divergence from schedule is a major source of higher capacity occupation which is ignored in capacity analysis performed using the schedule only.

The case study analysis identifies critical operations between Sargans and Chur which occur in realized train operation but stay hidden when only looking at the deterministic schedule. The expansion of the bottleneck concept by applying it to a critical path allows analysts to identify block sections that are most responsible for capacity occupation. Train operation must be adjusted at those locations to reduce capacity occupation on the whole line section. In the case study, large variations in running time, unrelated to the schedule critical path, result in higher occupied capacity via different critical operations. Thereby planners might be able to reduce capacity occupation of the operation by making small timetable adjustments considering realized operation. These improvements include adjusting the stopping pattern of a train to make operation more homogeneous or adjusting the speed profile and thereby the running time margin between stops.

The analysis of influencing factors presented identifies precise contributions of the train operation inside blocks and helps to identify which train dynamics and blocking time components render specific operations critical. In the case study, differences in stopping pattern and therefore acceleration, braking, and dwelling times are mostly responsible for lost time as part of capacity occupation. Moreover, since the block where the critical path has its longest vertical extent is most often a station without overtaking possibility, acceleration, braking and dwelling contribute most to capacity occupation. Differences in approach, running, and clearing time outside of stations do not play a large role (because all trains reach the maximum allowed speed). Specific knowledge of the effects of different braking and acceleration capacities can inform rolling stock acquisition helping identify where high-performance rolling stock is necessary to reduce capacity occupation (to improve stability or allow more trains) and where using cheaper, lower-performance rolling stock is sufficient.

The difference between capacity occupation in scheduled and realized operations is pervasive on all line sections we analyse. For example, on the line section between Sargans and Chur, realized capacity occupation lies roughly three percentage points above scheduled occupation. This represents an increase of capacity occupation greater than 10% on this line section. In short, planners must consider operational variations to plan robust schedules and not overestimate capacity. If infrastructure managers know the mean value and the distribution of the increase in capacity occupation between scheduled and realized operations, they can better define the amount of non-occupied capacity in the schedule needed to secure robust operations in real-life. In the case study of the Sargans-Chur line section, for example, the recommended value in UIC norm 406 of 75% maximum planned occupation could be seen as too conservative when only 3 percentage points more capacity is occupied in realized operation compared to the schedule. However, additional capacity occupation in realized operation compared to scheduled operation can be much higher if the realized operation is unstable. By analysing more case studies using a similar approach, the proposed maximum capacity occupation value of 75% in schedule could be assessed in its adequacy as a static recommendation.

Future research and practical applications should focus on a more precise recording of speed profiles and train locations to increase the data quality compared to reconstructed simulation data. Potential sources can be measured data points or high-resolution trajectories (e.g. based on online fusion algorithms). For instance, GPS-based localization can be linked to detailed signal-level data. Combining the method proposed in this paper with such data provides a detailed source for capacity planning by taking into account the variability of realized operation. Furthermore, the critical path concept of capacity assessment could be expanded to identify blockings of interacting trains that are not critical for capacity under the current operational regime but are interdependent because they cancel each other out. These blocks are especially relevant for expansion planning because they become critical immediately when the trajectory of one train changes.

Further approaches could discuss simplifying assumptions such as stochastic factors, which for instance limit correlation effects. Correlations are relevant, especially in serious disruptions. Approaches able to identify and model better correlations over spatial and temporal domains, like Büchel and Corman (2020) can identify to what extent the operations are vulnerable as a whole, and not as single elements. A similar direction is to analyse lines as a whole, avoiding splitting them into line sections. Currently the interdependencies between different line sections of a mainline railway, namely restrictions from the schedule or the requirements for continuous train paths along the whole line, are ignored by capacity analysis on line sections. The critical path approach presented in this research can provide a tool to tackle this shortcoming when developed further.

In this research, the safety system, the layout of signals, and the schedule are assumed fixed. For the future, it will be important to
investigate on how different, highly advanced operating schemes (see for instance the concepts reviewed in Nold and Corman, 2021) are differently sensitive to delays and result in more or less capacity occupation in planned operations. This research does show, however, that also variations in the vulnerability of this planned capacity occupation to realized operation must be considered and analysed. In principle, the concepts presented here are also applicable to these advanced technologies.

**Funding**

This work is partially supported by the Swiss National Science Foundation under Project 1481210/DADA, by the ETH Zürich Mobility Initiative, Project RailPower!, and by the ETH Zürich Foundation project number: 2023-FS-331 on Research in the area of Public Transport Design and Operations.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**


