Macroscopic Traffic Flow in Railway Systems – A Discussion of the Applicability of Fundamental Diagrams*

Norman Weik [a,1]
[a] German Aerospace Center (DLR),
Institute of Transportation Systems,
Lilienthalplatz 7, D-38108 Brunswick, Germany
[1] E-mail: norman.weik@dlr.de,
Phone: +49 (0) 531 295 3281

Abstract
Macroscopic traffic flow theory is widely used to analyze the performance and congestion in transportation systems and urban traffic. Currently, the main area of application is road traffic, where fundamental diagrams depicting the relationship between performance parameters such as traffic density, traffic flow and vehicle speed are used to represent traffic regimes and system settings. Research on the applicability for rail traffic systems is only starting to emerge and has mainly focused on mass transit systems and moving block signaling systems, so far. The current paper provides a theoretical analysis of traffic flow properties on unidirectional railway lines. Macroscopic fundamental diagrams are established and used to analyze similarities and differences of traffic flow behavior in railway systems. It is shown how fundamental diagrams allow to pinpoint flow regimes and different phases in train traffic, which can be useful information for system design and operation planning. Railway systems are studied both in a deterministic setting allowing to analyze peak efficiency and in a stochastic setting with perturbations.

Keywords
railways, traffic flow, capacity, fundamental diagram, congestion

1 Introduction

Macroscopic traffic flow theory allows to describe transport efficiency and to distinguish different traffic regimes and is widely used in modeling urban and road traffic. Fundamental diagrams depicting the interdependency between traffic density, traffic flow and vehicle speed provide a well-established means of representing physical characteristics of the traffic situation. In the railway context, the topic has recently seen increased interest as new signaling systems and moving block operations yield situations where train kinematics exhibit similarities to road traffic and traffic situations resemble phenomena previously discussed there. In particular, shockwaves and "jam" situations resulting from reactive behavior to preceding vehicles can be observed as a result of speed adaptations to keep safety margins between trains (Diaz de Rivera and Dick, 2021; Corman et al., 2019).

Whereas emerging signaling systems with moving block and very short train separations are natural applications for macroscopic transport theory, the situation is less clear for fixed-block systems. The same holds true for continuous train control in the presence of stationary way-side elements, such as switches, that need to be cleared a full braking distance ahead of an upcoming train for safety reasons. While delay propagation and effective train running times have been thoroughly studied in previous research, modeling procedures typically rely on models reflecting the discrete nature of dispatching or control decisions. A direct connection to transport theory and fundamental diagrams is rarely made, possibly due to the expectation that continuous traffic flow models provide limited insight into the complex nature of train dependencies and their implications on capacity and service quality.

*This is the Authors' Accepted Manuscript of the following article: N. Weik, Macroscopic Traffic Flow in Railway Systems – A Discussion of the Applicability of Fundamental Diagrams, Journal of Rail Transport Planning & Management 23, 100330, 2022, which has been published in final form at https://doi.org/10.1016/j.jrtpm.2022.100330 ©2022. This manuscript version is made available under the CC-BY-NC-ND 4.0 license, https://creativecommons.org/licenses/by-nc-nd/4.0/

Motivated by recent studies for metro systems (Farhi et al., 2017; Schanzenbächer et al., 2020; Cuniasse et al., 2015) and margin-operated trains (Corman et al., 2019; Diaz de Rivera and Dick, 2021), we analyze traffic phases in railway systems with fixed-block signaling systems using fundamental diagrams. Traffic situations are represented based on general system parameters and quality metrics including train arrival headway, train speed and line capacity (=traffic flow). Block lengths between 500m and 3km are considered – spanning "classic" fixed-block systems as well as shorter block segmentation partially resembling continuous train control systems.

The goal of this paper is two-fold:

- First, we describe fundamental diagrams for rail traffic on corridors and commuter lines and analyze the corresponding relationships including their implications on train operations. By providing a comprehensive overview of macroscopic traffic regimes in fixed-block railway systems on the global level, and summarizing and extending previous findings, a generalization is provided. We point out resemblances and differences to situations observed for vehicle traffic, for metro lines, and in recent moving block studies, discussing the capabilities and limitations of macroscopic traffic flow theory and the interplay of traffic flow metrics and railway operations.
- Second, we identify basic parameters to be read from macroscopic fundamental diagrams and how they relate to different infrastructure and traffic conditions. In particular, we provide examples how effects observed in macroscopic representations of traffic flow can indicate operational risks and pinpoint directions for improved system design and traffic control. Apart from a deterministic analysis of system parameters we study the influence of stochastic effects based on irregularities in train stopping behavior resulting from, e.g., differing passenger volume.

Our main focus are commuter rail corridors where relatively dense operations can be observed. As a result, continuity requirements for a meaningful traffic flow description are attained if long corridors and lines (w.r.t. mean train following distance) are considered (cf. Georliminis and Sun (2011)). In addition, system design and train control are particularly critical for this class of systems. Using a simulation approach we analyze different settings and discuss the interdependency of train operations and flow parameters.

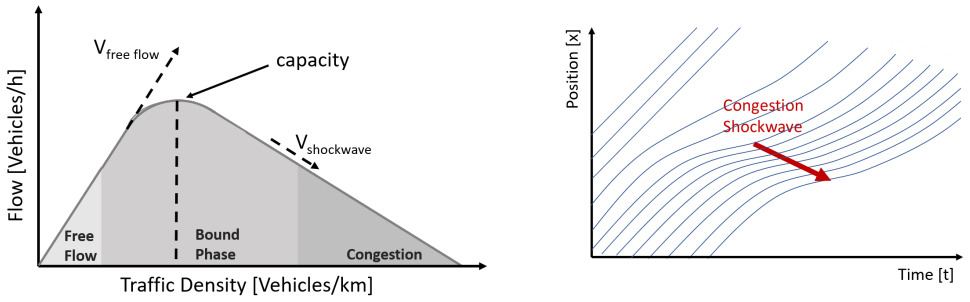
The presentation is structured as follows: We start by briefly reviewing the literature in Section 2, focusing on a short presentation of traffic flow theory for vehicular traffic, as well as related applications to railway systems. In Section 3 we present our methodological framework as well as modeling assumptions of the simulation approach used for system analysis. The settings for the analysis performed in this paper are described in detail in Section 4. Section 5 presents the results of our study of traffic phases in railway systems adopting both a deterministic and a stochastic view. The paper concludes with a short summary and an outlook to future research.

2 Related Work

In the following we briefly review the literature on traffic flow theory and macroscopic fundamental diagrams. We start by explaining the main concepts established for vehicle traffic. Subsequently, we review previous applications for railway systems and related concepts in delay propagation modeling.

2.1 Fundamental Diagram for Road Traffic

Assessing the efficiency and maximum flow capacity in transport systems has attracted the interest of researchers and transport planners alike for almost a century now (Knight, 1924; Wardrop, 1952). The topic remains a major task in transport planning, spanning different areas from network design to traffic control. By solving the kinematic equations associated with vehicle dynamics, macroscopic properties of the traffic flow on roadways can be obtained (cf., e.g., (Gerlough, 1976)). By graphically depicting macroscopic parameters such as vehicle density, speed and traffic flow in a *fundamental diagram*, basic traffic properties and different traffic regimes become visible. An illustration of the relationship between traffic density, flow and vehicle speed, as well as the individual vehicle trajectories are depicted in Figure 1.



(a) Fundamental diagram

(b) Vehicle trajectories

Figure 1: Fundamental diagram and individual vehicle trajectories for road traffic

In case of very low traffic load, all vehicles proceed at their designated speed without interaction, the system is in *free flow*. The slope of the curve in this region simply corresponds to the free-flow speed of vehicles and the flow of vehicles increases linearly in the number of vehicles (=vehicle density). Once a certain density is reached, vehicles start to interact and the system enters a *bound* traffic situation where vehicle kinematics are governed by vehicle interactions. Consequently the average vehicle speed starts to drop, while the overall traffic flow continues to increase up to a critical point, the *capacity*. Afterwards, traffic in the system remains afloat, but the traffic flow starts to drop as the vehicle speed drops superlinearily in the vehicle density. In this phase, congestion waves propagate, which result from vehicle interactions and the fact that drivers brake to maintain a safe separation to preceding vehicles. If the traffic density is further increased, the system ends up in a *traffic jam*. This overall dependency of traffic flow, speed, and density is typical for road traffic. A more detailed substructure and additional effects such as a capacity drop often encountered at the critical point can be observed. For a more detailed discussion and overview on different modeling and solution approaches see, e.g., (Gerlough, 1976).

2.2 Previous Discussions in the Railway Context

For railway systems, the global representation of traffic situations based on macroscopic fundamental diagrams has received relatively little consideration. Even though rail corridors in metro or long-distance networks theoretically seem to provide a similar study object, fundamental diagrams for rail traffic have only lately attracted increased interest.

In (Cuniasse et al., 2015), local fundamental diagrams for traffic in the Paris commuter rail network are analyzed. A dedicated capacity drop at the critical point marking the transition from the non-congested to the congested phase is found. Up to this point marking the maximum flow in the fundamental diagram, free-flow conditions are observed and the flow increases linearly in the traffic volume. At the critical train density, the traffic flow suddenly drops as a result of the speed control initiated by the signaling system once a train approaches a distant signal while the upcoming block has not been cleared by the previous train. In the congested phase, the flow is found to increase again up to the point where the train comes to a complete stop at the main signal and the traffic flow breaks down.

Sogin (2013) compares fundamental diagrams for road and rail traffic – noting a decisive difference depending on whether capacity limitations are considered for adjacent yards and segments or not. In addition, the effects of speed restrictions and multi-aspect signaling and speed control typical for American freight networks are discussed.

Diaz de Rivera and Dick (2021) extend these findings and study moving-block operation, focusing on main lines in long-distance networks. In particular, platooning of unidirectional trains is studied with the goal of increasing line capacity by close operation of successive trains. A saw-chain behavior and mild shockwave effects are found in the congested state in front of traffic obstacles such as, e.g., downhill slopes enforcing larger separation between vehicles due to increased braking distance. Changing line characteris-

tics are found to be critical for close operation of vehicles as the speed of lead vehicles needs to be decreased s.t. the original operating distance of the platoon can be re-established.

Corman et al. (2019) also discuss macroscopic fundamental diagrams for aggregate railway network analysis in signaling systems with moving block. For uniform, deterministic operation a fundamental diagram with a superlinear drop of the ow in the congested phase is found. For non-uniform, perturbed situations the ow is found to be non-decreasing. Even after reaching the critical density where trains start to interact, the ow continues to increase while the speed decreases. This is explained by the fact that the safety margin imposed by operation at safe braking distance between trains determines a minimum possible headway at the line entry point and consequently no complete jamming of the network can be observed. Furthermore, the mechanisms of train interactions enforced by the signaling system's constraints are studied and complex braking and accelerating behavior is found.

For metro systems, Farhi et al. (2017) establish a max-plus setting allowing to study train control in closed metro systems with a x number of vehicles and headway-dependent station dwell times representing increased passenger volume in case of irregular traf c. A fundamental diagram similar to the road traf c case with a free- ow phase, a plateau phase (saturation), as well as a linearly decreasing congestion phase is found. In (Schanzenächer et al., 2020), the model is extended to metro lines with two branches and additional traf c phases representing saturation of one or both branches are found.

Becker and Schreckenberg (2018) observe shockwave situations in densely operated railway systems with stochastic dwell times. A headway-dependent analysis of the effective train running time and ow is provided and compared to analytical calculations.

Corman and Henken (2022) perform a statistical analysis of rail traf c data from Switzerland. Macroscopic ow parameters are used to assess and distinguish different operating conditions and traf c states. In addition, the relationship between ow parameters and the observed arrival delays in train operations is investigated.

Szymula, Besinovic, and Nachtigall (2022) present a capacity planning approach for railway networks based on the combination of a Petri net scheme with scenario-based enumeration of different traf c scenarios. A network-wide macroscopic fundamental diagram is established from the minimal feasible occupation in harmonizing train circulations in the network. As a result, the effects of spatial and temporal correlations within the network can be quanti ed and the corresponding scheduled waiting times can be used to provide an indication on traf c quality.

2.3 Related Concepts for Railway Systems

Queuing models are another class of traf c ow models that have been applied to railway systems (Nießen, 2008; Wendler, 2007; Weik, Niebel, Nießen, 2016). Here, railway infrastructure is abstracted and aggregated to a single queue, where train service requirements are taken to be minimum headway times. Train interactions and occupation con icts within the infrastructure segment cannot be modeled – the service station is taken to be at the line entry point, where trains queue in case free running cannot be ensured, i.e. if train separation is smaller than minimum headway time. The performance of the infrastructure can then be assessed in an intrinsically stochastic approach based on waiting times resulting from irregular train arrival intervals and train-dependent minimum headways. In situations where the average arrival headway is smaller than the average minimum headway time the system becomes unstable and cannot be analyzed – even though, in reality, stable operations at reduced train speed might still be possible. To avoid this restriction of single queuing systems, Weik and Nießen (2020) describe a possible alternative based on a service line model, where different segments of the infrastructure are represented by individual queues such that the spillback of delays within a corridor can be accounted for.

Beyond queuing theory, which has a very general application scope in traf c ow analysis, many dedicated domain-speci c approaches for performance modeling and delay or runtime analysis have been discussed. For a general overview on capacity modeling, including simulation, timetable-based and -independent models, see (Abril et al., 2008; Pouryousef, Lautala, and White, 2015; Besinovic and Goverde, 2018).

Parametric relationships between traf c load and delays or running times have found widespread applications and are highly important for system robustness analysis and for incorporating service costs in transport optimization models. Borndoerfer et al. (2016) and Lieberherr and Pritscher (2012) make use of capacity-restraint functions for considering capacity utilization costs in the objective function of train routing problems. An exponential dependency on traf c density is assumed. The concept is borrowed from road traf c where comparable approaches are common. Detailed parametric models, typically taking into account widely accepted quality-determining factors such as train speed, traf c heterogeneity and initial delays, are also used for establishing relationships between traf c load and service quality (cf. UIC (2004)). This class of models is particularly popular in North America (Graham, 2008; Lai and Barkan, 2009), yet almost any infrastructure manager relies on comparable concepts in some regards. Gaining a better understanding of the interdependency of system and traf c parameters and the traf c situation observed in a network therefore continues to be important.

3 Model

In this section, we present our modeling and analysis approach for observing dependencies between traf c ow parameters that can be represented in fundamental diagrams. The analysis relies on a model that can be parametrized, such that fundamental diagrams for various railway system designs and traf c scenarios can be analyzed.

3.1 Setting

Infrastructure

To provide a setting for comparative analysis of transport properties in railway systems we de ne generic parametric railway systems, where system properties are tuned to account for different infrastructure situations and traf c regimes. The system under study consists of a unidirectional railway line consisting of stations (S) and line segments (L). The number of segments, the distance between stations, as well as the mean block length on line segments can be speci ed.

Two classes of systems are distinguished in the analysis:

- An open system allowing to study unidirectional traf c on a railway corridor. Here, trains are entering the system at the rst station at a pre-de ned frequency. Train headways (=arrival intervals) can either be deterministic or based on statistical input parameters. Upon arrival at the last station of the line, trains leave the systems and do not interact with other traf c any more. This class of systems allows to study traf c situations for unidirectional traf c on railway corridors. Results can be used to study capacity and throughput of heavily-used corridors, as well as to pinpoint possible bottlenecks and congestion risks.
- A closed system, where trains, upon reaching the last station, re-enter the rst station on their route and re-commence their service. This class of systems resembles closed circulation patterns for metro or commuter rail systems, where trains travel on speci c lines. In particular, the effects of rolling stock rotations and a limited number of train sets can be assessed. Here, train frequency and density is determined by the number of train sets operated on the line.

An illustration of the two model classes is provided in Figure 2.

Train Program and Train Dynamics

The modeling of trains is based on general train parameters such as length, speed, acceleration and braking capability, which are inherited from train classes or pattern trains. In the context of this paper, train dynamics is calculated using x acceleration and braking rates – a detailed consideration of traction and resistances is not incorporated. In addition, service-related parameters such as the train's route, scheduled stops and stopping times can be speci ed.

(a) Open system (b) Closed system

Figure 2: Classes of railway systems studied in this paper

Trains are generated at the first station along the route: For open systems, train frequency and headway are specified directly and trains are generated at the first station according to the specified headways. Headways may be perturbed and sampled from a probability distribution to account for running time variability and initial delays. In closed systems, the train frequency and headway is determined as a function of the number of trains in the system (N) and the operating period, i.e. the roundtrip time for a single free running train (rt_{free}):

$$t_{headway} = \frac{1}{f} = \frac{rt_{free}}{N}:$$

In both cases, trains are put to a waiting list in case no station track is available in the first station along the train route at the generation time and inserted at the earliest opportunity. For closed systems, the train order is kept upon arrival at the end of a circulation and arriving trains need to wait for preceding trains to depart before they can start their next service (here the number of trains N is the traffic control parameter).

3.2 Simulation Procedure

For the simulation of train operations, an event-based simulation is used which relies on a detailed representation of train infrastructure occupation and signaling requirements based on blocking times (see Figure 3). Whenever a train proceeds to the next infrastructure segment, a request for prolongation of the movement authority is generated whose timestamp corresponds to the instance the train reaches the distant signal within this segment or, respectively, the braking initiation point w.r.t. the end of the movement authority. In case the movement authority is not prolonged at the time the request is triggered, speed control and braking is initiated and the request is repeated with a certain frequency until the next infrastructure segment along the train's path becomes available and the train is allowed to proceed to the next segment along its route. In the following, the general request frequency is taken to be 1 s, however it can be parametrized to model different system and driver reaction times. For instance, the first repetition of the request could be prolonged and adapted as a function of a train's position to model situations where speed control can only be released after a certain time span, at a certain position, or once a defined target speed has been reached. In this paper this is considered in an approximate manner by prolonging the request cycle on line segments to 10 s.

In the context of this paper, the approach time is based on the braking initiation point w.r.t. the end of the block segments – resembling the situation in ETCS Level 2 and modern, block- and speed-optimized commuter rail systems. While a purely point-based operation based on distant signals, speed control initiated at the position of the distant signal, and the option for speed control release once the indication of the main signal can be determined could also be incorporated, it is not expected to yield fundamentally different results as the main physical parameters can also be modeled based on the braking initiation point and the request frequency.

3.3 Evaluation

For the evaluation, the following system parameters and service metrics are analyzed:

- train density on corridor (trains/km) – evaluated at train arrival instances,
- traffic flow (train arrivals/h),

Figure 3: Illustration of train simulation. A shockwave emerging from a bottleneck at the last station of the line can be seen.

- train running time on corridor (min),
- free (unimpeded) train running time on corridor (min),
- minimum headway time $h(t)$ – minimum temporal train separation at start of route to run unimpeded (in the absence of perturbations),
- average vehicle speed along route $v_{avg}(t)$ (km/h),
- trains in system $N(t)$ – measured at train arrival instances
- train frequency $f(t)$ (1/s) – train generation frequency (based on avg. input headway).

4 Case Study for Commuter Rail Systems with Fixed-Block Signalling

For the analysis of railway operation parameters and fundamental diagrams, a case study for commuter rail systems with homogeneous traf c and xed-block signalling is performed. The study comprises the following simulation settings:

- First, we make a distinction between a deterministic and a stochastic setting. In the deterministic setting, train frequency and station dwell times are x. This allows to study the system's performance at peak ef ciency. In the stochastic setting, station dwell times and train frequency at the beginning of the line are allowed to vary such that the effects of service variations and primary delays can be accounted for. Whereas frequency variations at the beginning of a train's route re ect initial delays, stations, where passenger board and disembark, typically form critical points on commuter rail lines. Train running times on line segments may also vary, e.g., due to different driving behavior, however, they generally exhibit less variation in homogeneous, densely operated commuter systems where operation is often supported by driver advisory systems or automated braking/acceleration systems.
- Second, we distinguish between a balanced and an imbalanced infrastructure scenario. In the balanced case, all stations are identical with equal train stopping times and train speeds. The imbalanced variant, by contrast, includes a built-in bottleneck at the last station on the line, where entry route, station and exit route are prolonged. This allows to study traf c behavior on the upstream corridor in front of the bottleneck.

In the deterministic setting, three line infrastructure designs with different system characteristics are studied. The two designs with a line length of about 50–60 km are taken as representatives of urban commuter lines with short inter-station distances of about 5 km, dwell times of 30 and 60 s (identical for all stations), and short average block lengths of 0.5 km and 1 km, respectively. The third line is taken to be longer (about 110 km) with larger mean station separation of about 15 km and average block length 3 km representing a regional connector line at the interface to a typical railway line, but still assuming homogeneous traf c. Dwell time for the connector line is taken to be slightly larger at 60 to 90 s. For the stochastic setting, the infrastructure variant with block length 1 km and dwell times of 60 s is studied.

Individual block lengths in the scenarios are allowed to vary by up to 10%, such that a more realistic non-laboratory line design is obtained. Block segmentation at stations is assumed to be capacity optimized. That means, for all stations, total station area length is assumed to 0.5 km in the balanced case. This comprises a platform area of length 167 m and entry/exit routes of the same length 167 m, each. For the imbalanced case with a built-in bottleneck at the end of the line, the station area of the last station is prolonged by a factor of 1.5.

An overview of all scenarios considered in the analysis is given in Table 1. l_{line} speci es the line length, l_{block} the mean block length, and t_{dwell} the station dwell time. mht denotes the minimum headway time, i.e. the minimum temporal separation of trains required upon corridor entry to ensure free- ow running unimpeded by other trains. Minimum headway times are calculated from blocking times for the respective scenario. In the scenarios for the regional line (l_{line} = 111.3 km) with station dwell time 60 s, the critical block is the rst block on the line after the second station. All other scenarios are station dominated, i.e. the critical block is located at stations where trains stop, and, consequently, dwell times are included.

Table 1: Infrastructure scenarios considered in the analysis

model setting	infrastructure	design	scenario	l_{line}	no. stations	l_{block}	t_{dwell}	mht
deterministic	balanced	open	1	111.3 km	7	3 km	90 s	170.8 s
			2	111.3 km	7	3 km	60 s	155.4 s
			3	56.4 km	11	1 km	60 s	140.8 s
			4	53.3 km	9	0.5 km	60 s	140.8 s
			5	53.3 km	9	0.5 km	30 s	110.8 s
		closed	6	111.3 km	7	3 km	90 s	170.8 s
			7	111.3 km	7	3 km	60 s	155.4 s
			8	56.4 km	11	1 km	60 s	140.8 s
			9	53.3 km	9	0.5 km	60 s	140.8 s
			10	53.3 km	9	0.5 km	30 s	110.8 s
	imbalanced	open	11	111.3 km	7	3 km	90 s	185.8 s
			12	111.3 km	7	3 km	60 s	155.8 s
			13	56.4 km	11	1 km	60 s	155.8 s
			14	53.3 km	9	0.5 km	60 s	155.8 s
			15	53.3 km	9	0.5 km	30 s	125.8 s
		closed	16	111.3 km	7	3 km	90 s	185.8 s
			17	111.3 km	7	3 km	60 s	155.8 s
			18	56.4 km	11	1 km	60 s	155.8 s
			19	53.3 km	9	0.5 km	60 s	155.8 s
			20	53.3 km	9	0.5 km	30 s	125.8 s
stochastic	balanced	open	21	56.4 km	11	1 km	60 s	140.8 s
		closed	22	56.4 km	11	1 km	60 s	140.8 s

In the analysis, trains are assumed to stop at every station and to use identical trainsets. Basic train characteristics including train length (l_{train}), train speed (v_{max}), acceleration (a_{acc}), deceleration (a_{dec}) used in the analysis are given in Table 2.

Table 2: Train characteristics used in the analysis

Train type	l_{train}	v_{max}	acc	dec
Commuter Train	100 m	100 km/h	0.5 m/s^2	-0.9 m/s^2

5 Results and Discussion

Based on the settings and scenarios described in the previous section we now present the results for the fundamental diagrams of rail traf c. To this end, the simulation is performed with different traf c loads. In the open system, this is achieved by varying the train arrival frequency, in the closed system the arrival headways are de ned by the number of trains operating on the line (cf. Section 3.1). For each data point a simulation of 12 operating hours has been performed. In the stochastic case, more than 150 realizations are contained for each infrastructure scenario and dwell times and arrival headways are sampled from normal distributions with coef cient of variation $v_X := \frac{\sigma}{\mu}$, where μ is the mean and σ the standard deviation of the distribution. In order to avoid unrealistically short (and even negative) station dwell times and arrival intervals, the distribution is cut at $\mu \pm 50\%$. For each 12-h-simulation run, the statistical time-averaged metrics described in Section 3.3 are calculated to assess the long-term behavior and performance of the system in terms of macroscopic traf c ow properties such as (spatial) train density and average vehicle speed or running time.

5.1 Deterministic Analysis for Balanced Infrastructure

Figure 4 depicts the results for balanced infrastructure with identical stopping times at all stations. The left-hand side shows the global, corridor-based macroscopic fundamental diagrams in the open system, the right hand side provides the corresponding gures for the closed system. In the rst row, the traf c ow is depicted as a function of the train arrival frequency or, respectively, the closely related number of trains in the closed system. The second row shows the average train speed over the arrival frequency. Finally, the last row depicts the ”classic” road traf c representation of traf c ow over vehicle density on the lines. Traf c ow is measured as the ow of arriving vehicles at the end of the corridor/last station, train density refers to the average density of vehicles on the entire line and average speed is the average vehicle speed on the line – not accounting for potential queuing upon line entry at the beginning of the corridor. In all diagrams, the results for the ve infrastructure scenarios with different block length and stopping time according to Table 1 are depicted.

Open System

In the open system, the traf c ow is found to increase linearly in the train arrival frequency up to a critical point, where it suddenly drops and saturates at a lower, constant ow (cf. Figure 4a). For Scenario 2, only a very small capacity drop at the critical point can be distinguished before the system settles at a constant ow. Scenario 1 with the largest minimum headway time has the lowest capacity, Scenario 5 with the smallest, the highest. Scenarios 3 and 4 have almost identical ow for all train arrival frequencies. The two scenarios belong to different line characteristics, but share the same minimum headway times and station dwell times.

The nding of the sharp capacity drop at a critical traf c load resembles the observation made in (Cuniasse et al., 2015) where this point demarks the start of train interactions and the end of the free- ow phase if trains encounter a restricting distant signal. Beyond that point, the system adopts a new equilibrium with smaller throughput, which is based on the new congested blocking times of trains in the system. It is, however, remarkable that in our ndings the capacity drop in all scenarios is located at a train arrival headway several seconds below the respective minimum headways. In Scenarios 3 and 4, for instance, the phase transition occurs at an arrival frequency of 28.1 trains per hour (corresponding to an arrival headway of 128 s) and, hence, almost 13 s below minimum headways. Considering the average vehicle speed on the corridor depicted in Figure 4c two drops of velocity can be identi ed in all scenarios but Scenario 2: One

(a) Flow over arrival frequency (open)

(b) Flow over no. of trains (closed)

(c) Average speed over arrival frequency (open)

(d) Average speed over no. of trains (closed)

(e) Flow over train density (open)

(f) Flow over train density (closed)

Figure 4: Fundamental diagrams for infrastructure with balanced design

(a) Headway 130 s, stable flow (b) Headway 127 s, instable flow (c) Headway 125 s, instable flow

Figure 5: Train trajectories in Scenario 3 in the vicinity of the critical point. Stable flow (left), instable flow at the critical point (middle) where the fourth station acts as the bottleneck, and instable flow in the congestion phase beyond the critical region (right) with the bottleneck forming at the first station acting.

significant drop at the position of the capacity drop in Figure 4a, and, though small in comparison, another one further left. The latter coincides with the minimum headway times on the line (at an arrival headway of 140 s, i.e. an arrival frequency of 25.7 trains/h in Scenarios 3 and 4, for instance). Even though trains start to interact at this point, the system seems to remain stable and is still able to buffer small delays originating from the bottleneck. Hence the further increase of traffic flow even though train headways fall below minimum headways.

The train operating graphs in Figure 5 based on Scenario 3 with l_{block} = 1 km illustrate the system behavior in the vicinity of the critical point. To the left, at a headway of 130 s, the system still exhibits steady flow, whereas at a headway of 125 s (right) congestion has formed upstream of the first station. This is to be expected as all stations are identical (in terms of dwell time and route length), s.t. the first station on the line should act as the bottleneck in the congested phase in case the critical blocks are located at stations. At the critical point exactly, the situation is found to be less clear. Figure 5b shows that an intermediate station (the forth) seems to be the bottleneck in this case – which also reflects macroscopically in Figures 4c and 4e, where a single overshooting point is detected at the location of the phase transition, exactly. The cause probably is a random effect resulting from the exact block lengths in the forefront and aftermath of stations which influences delay propagation and points out this station. As the corridor to reach this system state seems to be very narrow it will probably not be stable and hard to reach in practice.

Unlike for road traffic, no complete breakdown of train traffic will occur in the congested phase. In this phase, train operation is still possible on the corridor – under reduced speed enforced by the signaling system by means of speed control and block access rules. System throughput in the congested regime is thus controlled and determined by the worst operational train blocking time at or upstream of the bottleneck. Behind the bottleneck, an increased train headway (depletion zone) is observed as throughput is delimited by the bottleneck (cf. Figure 5). Initial train feeding frequency to the corridor adjusts to this new traffic state once the congestion density wave has propagated back to the line entry point – a finding also reported in (Corman et al., 2019) for perturbed operations. As a result, a new stationary congested regime with reduced traffic density is observed, which can also be seen from the Figure 4e, where the train density increases at the critical point and then decreases again, once the bottleneck at the first station has delimited downstream flow in the congested phase, s.t. to the right of the critical point, flow and train density converge to a single point.

Closed System

In the closed system, system capacity also generally reflects the different minimum headways. However, a richer substructure with up to four different traffic regimes can be distinguished in the fundamental diagram:

The first one is the free-flow phase. Whereas the slope of the fundamental diagram is similar for Scenarios 6-8, the flow in Scenarios 9 and 10 increases more slowly in Figure 4b as a function of the number of operating vehicles. This can simply be explained by the different line length, s.t. the same number of vehicles operating on the line corresponds to different vehicle densities. In Figure 4f, where traffic flow is depicted as a function of the vehicle density, it can be seen that the slope of the diagram in the free-flow phase and hence the average vehicle speed (cf. the discussion in Section 2.1) is comparable.

At the end of the free-flow phase, a small capacity dip is observed in all scenarios but Scenario 7 (purple) – the scenario where the decisive block for minimum headways is on a line segment and not within a station (a similar behavior was also seen and previously discussed for the open system). The average speed drops, as can be seen in Figure 4d and a new phase is attained. In this second phase, traffic flow can be seen to slowly increase again in Scenarios 9 and 10. If the train number exceeds 30 (Scenario 10) or 27 (Scenario 9) the flow does not increase further and the system reaches a plateau phase. For the other scenarios this phase seems to be attained immediately after the free flow phase. The length of the plateau phase reflects the different infrastructure scenarios and is longest for the scenarios with block length 500m and shortest for those with block length 3km. If traffic load is increased further, another decisive capacity drop occurs at the end of plateau phase and the traffic flow starts a continuous decline. Phases (1) free flow, (3) plateau and (4) breakdown have also been observed in (Farhi et al., 2017; Schanzenbächer et al., 2020) for closed-line metro systems, the second phase is not reported there.

Phases (2) to (4) can be explained by the following system characteristics: Starting in free flow and increasing the traffic density, trains interact and delays start to form form once train headways reach minimum headway times. As a consequence, train speed and traffic flow drops. In this phase, a congestion wave propagates back from the bottleneck and causes the system to adopt a new equilibrium at reduced capacity. For balanced infrastructure, the first station on the line will be the position where interactions start, and downstream flow from that point on, will be reduced. This reduction of downstream flow results in an increase of train headways in this area, and, as the system is closed, also at the initial upstream station once trains re-arrive there after completing their round-trip journey. As a result, a readjusted train circulation pattern with increased trip running time is obtained.

The saw-chain behavior of the transition from phase (1) to (2) has also been observed in (Diaz de Rivera and Dick, 2021) and (Cuniasse et al., 2015). There, it was found that this behavior is due to the block system where train braking and speed control is initiated once train density reaches the point where succeeding trains are given a braking request when approaching the distant signal or braking initiation point. The system is more efficient and traffic flow increases again if the train distance is further decreased in this new train-braking regime, as train speed drops sublinearly. This sublinear behavior of average train speed on the line can also clearly be seen for all scenarios in Figure 4d. The (max,+)-formulation presented in (Farhi et al., 2017; Schanzenbächer et al., 2020) does not model train speeds, which is probably the reason why this phase is not observed there.

Further increasing traffic volume, the flow remains at a constant value and a plateau phase (3) is attained. This phase is reached once the train number is so high that it affects the circulation of rolling stock on the line. Starting from this point, trains cannot re-commence their service upon reaching the end of their journey as they need to wait for previous trains to start their service and leave the first station, as a result, speed continues to drop. Still, the system remains stable even in the plateau phase once the congestion has propagated and operation speed is further reduced. Train operation in this regime is not efficient, though, as the effective transport capacity on the corridor is constant in spite of the increase of vehicle number.

This behavior holds until a tipping point is reached and traffic starts to break down (Phase (4)). At vehicle densities beyond this point, interactions between vehicles in the closed system become so strong that large-scale blocking occurs and traffic flow and average train speed drop dramatically. The position of this tipping point is related to block structure, which can clearly be seen from Figure 4b. Ultimately, the system ends up in a complete lockdown if additional vehicles are inserted.

5.2 Deterministic Analysis for Imbalanced Infrastructure

The results for the imbalanced system, where the last station on the line is designed to be the capacity-limiting factor by increasing station length by a factor of 1.5, are depicted in Figure 6. Corresponding infrastructure scenarios 10-15 (open) and 16-20 (closed) are given in Table 1.

Open System

In the open system, the free- ow phase and the congested phase are similar to the balanced case with a sharp capacity drop at the critical point marking the start of interactions yielding backpropagation of delays. The observed ow is lower in all scenarios but Scenario 13, re ecting the changes of the minimum headways in Table 1. For this speci c scenario, the prolongation of the nal station has resulted in a blocking time there which is only slightly higher than the one at the rst block section on the second line segment – hence the very similar ow as in the balanced case. Unlike the balanced case, Scenarios 11-13 now all exhibit the same traf c ow in the congested region – which can be explained by the fact that these scenarios share the same minimum headway times with the nal station (with a dwell time of 60 s featuring the longest block occupation time.

While the traf c ow behavior resembles the balanced case, the transition region after the capacity drop differs as can be seen from the average train speed in Figure 6c. In this regime, a gradual drop of speed associated with a gradual increase of train density on the line (cf. Figure 6e) is observed, until, at some point, the speed goes up and the density reduces again and stays stable in the nal, congested phase. This behavior does not show in the balanced case, where a sharp drop of the average speed at the critical point (possibly including an overshoot) has been observed.

The reason for the more gradual decline in speed has to do with the location of the bottleneck. As the imbalanced case has a bottleneck towards the end of the line, the resulting congestion spillback requires some time to reach the beginning of the line where vehicle entry headways are adjusted. In Figure 7, the time-distance graphs in the area of this traf c phase and around the capacity drop for Scenario 13 are depicted. At a train arrival headway of 145 s (left), the system is still at free ow. Then, once the capacity drop has occurred (headway 130 s, middle), delays start to backpropagate from the last station demarking the bottleneck. This marks a traf c phase where system behavior is governed by the bottleneck at the last station resulting in a gradual ll-up of the system with trains. It can also be seen that propagation of delays is relatively slow as a result of the relatively low train density (as compared to road traf c), such that the gradual decline of train speed and increase of density is a result of the non-stationary behavior for the limited time horizon of 12 h. The higher the train arrival rate the faster the backpropagation of the density wave and the ll-up of the system.

If train arrival headways are further decreased beyond the capacity drop in the imbalanced system, an additional congestion wave originating at the rst station (headway 125 s, Figure 7c) can be seen like in the balanced system. As a result, the traf c ow downstream of this station is already limited and the strength and propagation speed of the congestion wave originating from the bottleneck at the end of the line are reduced. Macroscopically, this means that the train density on the line will decrease and the vehicle speed will increase again once the transition region is passed.

In the long run, the congestion caused by last station may eventually govern the entire line (this situation is observed in Figures 7b and 7c). Alternatively, the last station may become uncritical as the ow-delimiting effect of an upstream station dominates – a situation which is observed in Scenario 12, where the bottleneck at the nal station is weak in comparison to the blocking times at upstream line segments. As the rst block section on the second line segment is almost as critical, trains are blocked upon exit from the preceding station, which, subsequently, becomes the decisive one. This re ects in the macroscopic ow parameters in Figure 7c and 7e, where speed and density converge to a constant value after the capacity drop without showing any sign of another phase transition. Depending on the system design and the nature of interactions, different phenomena including induced or masked bottlenecks, where the downstream bottleneck becomes uncritical once trains start to interact at an earlier location, are possible in this phase depending on how the system react to train interactions, locally. As a result, the transient behavior re ected

(a) Flow over arrival frequency (open)

(b) Flow over no. of trains (closed)

(c) Average speed over arrival frequency (open)

(d) Average speed over no. of trains (closed)

(e) Flow over train density (open)

(f) Flow over train density (closed)

Figure 6: Fundamental diagrams for infrastructure with bottleneck at the end of the line