Table of contents

Summary

Public transport (PT) plays an important role in urban mobility and accessibility, with reliability, particularly punctuality, being a key determinant of passenger satisfaction and long-term ridership. Traditional metrics, such as On-Time Performance (OTP), often focus on vehicle performance and tend to overlook the passenger experience. This report explores alternative, punctuality metrics with more focus on the passenger perspective that more accurately reflect passengers' actual travel experiences, i.e., passenger-centric punctuality metrics.

The research begins with a qualitative review of both traditional and passenger-centric metrics, providing insights into variations of OTP and other measures. These metrics offer a more in-depth understanding of service reliability by considering real-time passenger demand and travel patterns. A quantitative case study on Stockholm's commuter rail network, using traffic and passenger demand data, highlights differences between vehicle-centric and passenger-centric measures, particularly during peak and off-peak times.

Key findings:

- Various passenger-centric metrics are applicable to monitor PT service reliability more accurately by accounting for passenger demand and travel experiences across different modes and time periods.
- The reviewed metrics vary in terms of relevance, accuracy, and ease of implementation for monitoring passenger punctuality in PT systems.
- One major challenge in implementing these metrics is the limited availability and quality of passenger demand data, as well as the technical complexity involved in integrating automatic data collection technologies with existing systems.
- Passenger-centric metrics can enable better monitoring of passenger punctuality and better operational management, especially in spatio-temporal areas where demand variations are significant.

Adopting passenger-centric metrics is essential for improving PT performance monitoring. These metrics can provide valuable insights that traditional vehicle-centric approaches may overlook, particularly in understanding the impact of passenger demand variations in space and time. By combining passenger-centric and vehicle-centric metrics, PT stakeholders can make more informed, data-driven decisions that enhance service performance, operational efficiency, and ultimately, passenger satisfaction. However, further research is needed to assess the benefits (and costs) of combining both perspectives, as such a comprehensive approach could enable more targeted improvements, particularly in areas with high passenger delay prevalence.

Sammanfattning

Kollektivtrafiken spelar en viktig roll för urban mobilitet och tillgänglighet, där tillförlitlighet, särskilt punktlighet, är en nyckelfaktor för att säkerställa passagerarnöjdhet och långsiktigt resande. Traditionella mått, såsom "On-Time Performance" (OTP), fokuserar ofta på fordonens prestanda och tenderar att förbise passagerarnas upplevelser. Denna rapport utforskar alternativa, passagerarcentrerade punktlighetsmått som mer exakt speglar passagerarnas faktiska reseupplevelser, det vill säga passagerarcentrerade punktlighetsmått.

Forskningen inleds med en kvalitativ genomgång av både traditionella och passagerarcentrerade mått och ger insikter i olika varianter av OTP och andra mått. Dessa mått ger en djupare förståelse för servicepålitlighet genom att ta hänsyn till passagerares realtidsbehov och resmönster. En kvantitativ fallstudie av Stockholms pendeltågsnät, baserad på trafik- och passagerardata, belyser skillnaderna mellan fordoncentrerade och passagerarcentrerade mått, särskilt under hög- och lågtrafik.

Viktiga slutsatser:

- Det finns olika passagerarcentrerade mått som kan användas för att mer noggrant mäta kollektivtrafikens tillförlitlighet genom att ta hänsyn till passagerarnas efterfrågan och reseupplevelser över olika transportmedel och tidsperioder.
- De granskade måtten varierar i fråga om relevans, noggrannhet och implementeringsmöjligheter för att övervaka passagerarnas punktlighet i kollektivtrafiksystem.
- En viktig utmaning vid implementeringen av dessa mått är begränsad tillgång till och kvalitet på passagerardata, samt den tekniska komplexiteten som krävs för att integrera dem med olika befintliga system.
- Passagerarcentrerade mått möjliggör bättre övervakning av passagerarpunktlighet och förbättrad operativ styrning, särskilt i spatio-temporala områden där efterfrågevariationerna är betydande.

Att anta passagerarcentrerade mått är avgörande för att förbättra uppföljningen av kollektivtrafikens prestanda. Dessa mått kan ge värdefulla insikter som de traditionella fordoncentrerade metoderna ofta förbiser, särskilt när det gäller att förstå effekterna av variationer i passagerarnas efterfrågan över tid och rum. Genom att kombinera passagerarcentrerade och fordoncentrerade mått kan aktörer inom kollektivtrafiken fatta mer informerade, datadrivna beslut som förbättrar serviceprestanda, driftseffektivitet och slutligen passagerarnöjdheten. Däremot behövs ytterligare forskning för att utvärdera fördelarna och kostnaderna med ett mer omfattande tillvägagångssätt som kombinerar båda perspektiven, då detta skulle kunna möjliggöra mer riktade förbättringar, särskilt i områden där passagerarförseningar är som mest frekventa.

1. Introduction

Quality public transport (PT) systems play a vital role in providing sustainable mobility and accessibility for urban areas. The quality of PT services affects the overall attractiveness of PT systems and is therefore important for attracting new passengers and for keeping existing ones. Reliability is one of the important quality aspects of PT services, as it can significantly affect passenger experiences and user satisfaction in the long term (Karlsson et al., 2011). However, the reliability of PT services is often challenged by various factors, such as traffic congestion, operational disruptions, and demand fluctuations (Wei et al., 2024).

One of the most visible components of reliability in PT systems is punctuality; as it directly affects the travel experience of passengers (travel time, waiting time, convenience), and long-term satisfaction. PT service punctuality is shown to have a strong effect on how satisfied passengers are and how they view the overall quality of PT services (Friman, 2004).

1.1. Background

To monitor and measure the punctuality of PT services, traditional metrics have mainly focused on the punctuality of vehicles instead of passengers. For instance, to monitor the punctuality in Swedish railways, the infrastructure manager, or Trafikverket (2020), uses the percentage of trains arriving, at the final station, within 5 minutes (commonly noted *RT+5*). Such vehicle-focused metrics are the basis for many analyses and policy recommendations. Although beneficial from an operator's perspective, these punctuality metrics often neglect the passengers' viewpoint and therefore omit factors such as variations in travel demand (den Heijer, 2018), passengers' valuation of travel time (Barabino et al., 2015) and waiting time (Ait Ali et al., 2022).

A number of studies have shown a tendency to move to passenger-oriented metrics because traditional measures are focused on the operators' perspective (supply side) and may not match passengers' travel patterns (Bagherian et al., 2016). Therefore, policies and strategies that aim to improve the quality of PT services may not have the intended effects. Hence, such a move toward passenger-oriented metrics seems essential.

This shift to the demand/passenger side is further driven by the emergence of new technologies and valuable sources of new data (Pelletier et al., 2011), such as Automatic Vehicle Location (AVL), which includes traffic control data or GPS data on vehicles to monitor real-time locations and movements. Automatic Fare Collection (AFC) systems use technologies like smart cards or mobile payment applications to track boarding patterns and fare transactions. Automatic Passenger Counting (APC) systems are based on sensors, such as infrared sensors and captors on doors to count passengers entering/exiting the vehicle. These technologies collectively provide a richer and more

precise understanding of passenger demand and operational dynamics in public transport (Ghofrani et al., 2018).

1.2. Problem statement

Several studies have shown that punctuality from the passenger side, i.e., passenger punctuality, is generally lower than traditional vehicle punctuality. Vanhanen and Kurri (2005) study quality factors in Helsinki's PT service and show that the passengers' perceived quality, including reliability and hence punctuality, may differ quite significantly from the technical service level indicators employed by the operator's planners. More recently, Nelldal et al. (2019) show, using average passenger loads, that punctuality levels obtained using *RT+5* are higher (up to 8.7%) compared to passenger-weighted punctuality. The authors highlight therefore the importance of weighting train punctuality based on the number of passengers to obtain a more accurate measure of passenger punctuality. A recent report by Transportstyrelsen (2023) states that 60% of train passengers think that trains are punctual whereas statistics indicate that train punctuality is as high as 90% (using *RT+5* metric). Moreover, surveys on the passengers' satisfaction with service punctuality indicate that there is a gap between what the statistics show and what the passengers experience.

One reason for this gap is that passengers tend to perceive punctuality as worse than it is because they tend to remember more the unpunctual/delayed trips rather than the punctual ones, i.e., risk aversion (Nielsen, 2000). Another reason is that delays tend to be more common/longer during peak periods when more passengers travel in the PT system and therefore traditional vehicle punctuality measures underestimate passenger delays (Parbo et al., 2016). Another major drawback of such traditional measures is not fully capturing the effect of delays on passenger travel times and transfers, e.g., a small delay on a vehicle can lead to many passengers missing their connections and result in longer travel and waiting times. Hence, traditional vehicle-centric punctuality does not generally reflect passenger delays in PT systems. There is therefore a gap between these traditional metrics and the actual passenger experience, which requires more research exploring alternative metrics that consider and better capture punctuality in PT systems from the passengers' perspective.

1.3. Aims and objectives

The study aims to explore alternative approaches to monitor punctuality as an important reliability component and the consequence of quality PT services. The focus is on how punctuality can be measured and better monitored from a passenger perspective. The goal is to provide recommendations for improved PT management and lay the foundation for future research aimed at analyzing and enhancing the monitoring of punctuality in PT systems from a more passenger-focused perspective.

Based on an assessment of possible metrics and an investigation of their limitations, challenges, and potentials, the main objective of this work is to provide initial research

2. Passenger punctuality in public transport

PT systems are expected to transport passengers to their destinations with minimal delays and disruptions. Thus, punctuality is often used as an indicator of the performance and reliability of these systems. However, measuring and monitoring punctuality is complex, as it depends on multiple factors, including network characteristics, demand patterns, service design, traffic conditions, and operational management. Therefore, it is useful to review and understand some existing research and practices that are related to measuring and monitoring punctuality and to identify metrics for passenger punctuality that can be in PT systems.

2.1. Punctuality as an indicator of service reliability

In transport systems, punctuality is a fundamental indicator of service reliability, alongside regularity, particularly for passenger transport like PT systems (Gittens and Shalaby, 2015). From an operational point of view, Rudnicki (1997) defines punctuality as any deviation, expressed in absolute or relative units, from the timetable, i.e., schedule adherence. Punctuality reflects therefore the system's ability to transport passengers to their destinations on time. In practice, it is often measured using OTP, i.e., the proportion of trips meeting specified punctuality thresholds, such as being within 5 minutes of scheduled arrival time, also known as *Right-on-Time* or *RT+5* (Rudnicki, 1997). The thresholds are usually lower for short-distance PT services than for inter-city and long-distance services, e.g., thresholds for commuter services range from 3 to 6 minutes (Denti and Burroni, 2023).

Different perspectives can be used to evaluate PT service reliability (Zhao et al., 2013). The Transit Capacity and Quality of Service Manual, or TCQSM (2013), states that service quality and reliability are measured by how PT services are perceived by passengers. PT agencies or operators may have a different view, focusing more on how well PT services are operated rather than the passengers' experience.

Extensive research has explored punctuality across various dimensions, including definitions, metrics, costs/benefits, causes, and effects, highlighting its critical importance to passengers, operators, and PT agencies. Two main punctuality perspectives are highlighted in the literature, namely the vehicle and passenger perspective, see Figure 2 by van Oort (2016) for an illustration of the main determinants in both perspectives. Unpunctual operators face increased costs due to schedule recovery challenges, often requiring extra vehicles to maintain the promised level of service, i.e., travel time, and/or frequency (TCQSM, 2013). For passengers, unpunctuality can lead to longer travel and waiting times and uncomfortable trips, e.g., crowding onboard or missing connections,

especially for passengers following fixed departure times, e.g., in the case of infrequent services with longer headways (Bagherian et al., 2016). Improved punctuality positively impacts passenger perceptions and can boost PT ridership (van Loon et al., 2011).

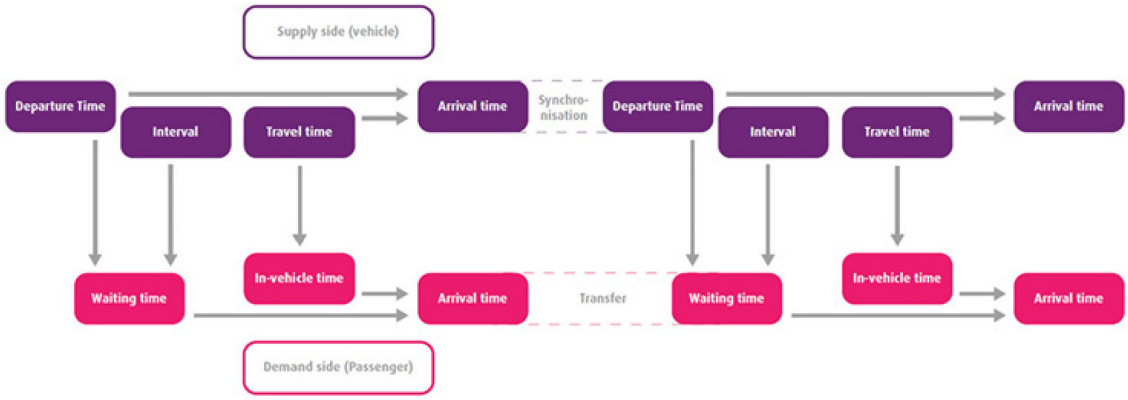


Figure 2. Passenger and vehicle trip determinants of service punctuality (van Oort, 2016).

To differentiate punctuality measurement levels in PT systems, Danaher et al. (2020) distinguish between system, route, trip, and stop level, each offering specific insights into punctuality analysis. Often in practice, instead of analysing the performance on all routes, which is expensive/unrealistic, a sample of main routes is chosen for larger areas of ridership, demographics, and geography in a PT system (Cramer et al., 2009). Additionally, various statistical techniques are employed to analyse punctuality, e.g., percentage analysis, time unit comparisons, normalization, and variability indices (Danaher et al., 2020).

Moreover, certain types of metrics are introduced to account for, e.g., time-of-day and day-to-day variability while reflecting both passenger and transit agency perspectives (Chan, 2007). For instance, distributions of journey time (total travel, in-vehicle, and wait times), with metrics such as mean, *coefficient of variation* (CoV), and the percentage of observations exceeding specific thresholds. For high-frequency, e.g., peak-hours PT services, schedule adherence is measured at specific stops of a route by assessing the average deviation from the schedule, its coefficient of variation, and the percentage of deviating arrivals. Headway distributions can also be measured, including their mean, coefficient of variation, and percentages of headways with certain thresholds.

The use of some or a combination of these techniques, from a passenger or operator perspective, can allow for more comprehensive comparisons of punctuality levels within and across different parts of the PT system. **Table 1** provides some examples of metrics, their type and level. Some are expressed in time units, e.g., delay average, or metrics in the form of percentages or distributions, e.g., OTP, which are easier to understand but do not capture the actual variations in punctuality. Variability metrics, e.g., variance, standard deviation, and CoV, quantify the extent of variation in relevant aspects of the passenger trip, e.g., waiting and travel times. For instance, CoV is unitless and can be used for comparisons between different services, e.g., routes/lines with different properties. Indices based on high percentile values indicate the limit of common deviations from the mean, while ignoring rare extremes and outliers, and can be used to

compare services with different mean values. Other classifications include central tendency versus extreme risk indicators, as well as binary/discrete versus continuous (Blayac and Stéphan, 2021).

Table 1. Examples of punctuality metrics, their corresponding type, and level.

Type	Example	Level
Units of time	An average delay at a given stop/station is 5 minutes.	Stop
Percentages/distributions	The percentage of trips on time (arrival within 5 minutes to destination) from an origin is 90%.	Trip
Variability measures	A route, with a 20-minute average running time and a standard deviation of 3 minutes, has 0.15 as a CoV.	Route
Indices/percentiles	The 95th highest running time in the bus network was only 39 minutes, i.e., an index of 1.3 when divided by the average running time.	System

2.2. Commonly used punctuality metrics

Choosing an appropriate metric is important for evaluating the punctuality of PT services. As mentioned earlier, one commonly used metric is OTP, which calculates the percentage of "on-time" arrivals, often to the final stop. "On-time" is generally within a specified threshold or delay tolerance window compared to the scheduled time (Rudnicki, 1997, Danaher et al., 2020). Given a delay threshold τ, OTP can be calculated across all stops S using equation (1).

$$OTP(\tau, S) = \frac{\sum_{i \in S} V_i(\tau)}{\sum_{i \in S} V_i^{tot}}, \quad (1)$$

where $V_i(\tau)$ is the number of operated vehicles arriving within a delay threshold τ to stop i, whereas V_i^{tot} is the total number of operated vehicles (often excluding cancellations). Instead of considering all stops S, OTP is often calculated at terminal stations or stops.

2.2.1. Variants of OTP

While OTP, as defined in equation (1), provides a straightforward measure of punctuality, it has its limitations, particularly in its binary nature, which overlooks cancellation and/or the magnitude of deviations from the schedule (Barabino et al., 2015). Delay tolerance, or punctuality windows, are critical factors in determining OTP scores and ultimately, passenger punctuality. Different countries and operators employ varying criteria, emphasizing the importance of understanding how these criteria impact service quality (TCQSM, 2013, Blayac and Stéphan, 2021). For instance, the British rail infrastructure manager (IM) relies on OTP-variants such as the "*time-to-X*" measure, i.e., the percentage of recorded stations/stops (not vehicles) where trains arrive within a threshold X (varying between 1 and 30 minutes), alongside the official 1-min threshold OTP (NetworkRail, 2017).

The British IM also measures OTP at each passenger station to capture the proportion of trains arriving on time at intermediate stops (besides terminal ones). This can better capture passengers' real experiences, especially early departures and longer waiting times (Mishalani et al., 2006, Denti and Burroni, 2023). In the case of long-headways services,

Zhao et al. (2013) mention the use of a timetable-based metric similar to OTP, i.e., *En-route Schedule Adherence* (ESA), but assesses schedule adherence (often between -1 and $+5$ minutes) along multiple en-route time points. ESA has been used to monitor punctuality in the PT system of New York, and is more suitable when passengers rely on published schedules, e.g., low-frequency PT services during off-peak hours (Cramer et al., 2009).

To monitor train service punctuality, the Swedish IM, or Trafikverket, uses an alternative measure, i.e., *delay increment* (DI), which calculates deviations (of at least 3 minutes) between a train's actual passing time and its scheduled time at consecutive points along the route, see **Figure 3** for an illustration by Joborn and Ranjbar (2022). Each disturbance leading to a non-zero DI is categorized by cause for monitoring infrastructure failures and effects on train punctuality. DI is used to implement reliability improvement plans and for accountability with train operators using performance regimes (Joborn and Ranjbar, 2022). The same authors proposed the concept of *delay contribution* (DC) as an improvement to DI to better identify the critical disturbances that have more effects on the train delays, e.g., officially measured using OTP at the final station.

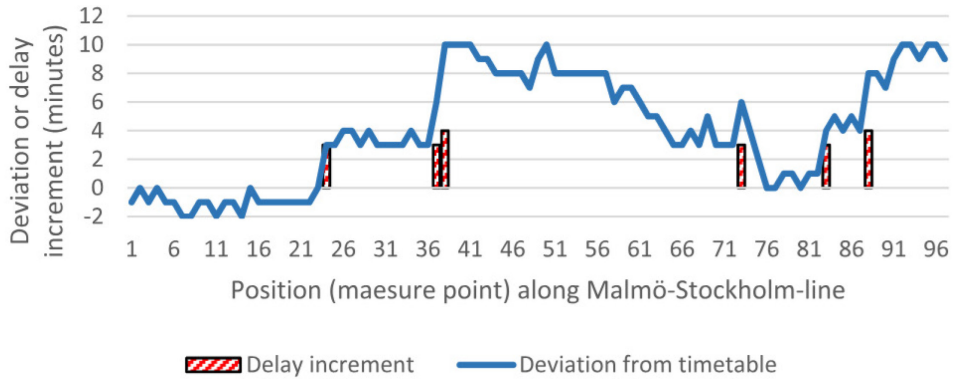


Figure 3. Illustration of delay increments (DI) used to monitor train service performance in Sweden, image by Joborn and Ranjbar (2022).

To account for cancelled services, variants of OTP have been introduced such as the *Combined Performance Measure* (CPM) which considers both delayed and cancelled departures in the total scheduled departures, i.e., V_s^{tot} includes cancelled departures in equation (1) leading to lower scores than OTP (Parbo et al., 2016). Until recently, CPM was used by Trafa (2023) for reporting all official punctuality statistics of rail transport in Sweden. To report cancellations in British rail services, the IM uses cancellation scores at all recorded station stops to complement the "time-to-X" metric (NetworkRail, 2017).

From a demand perspective, passengers think about punctuality, and reliability in general, in terms of the overall waiting and travel time. Unlike operators (from a supply perspective) which focus on how well services follow OTP standards, passengers are more affected, in their departure time decisions, by waiting and running times and their variations (Diab et al., 2015). Hence, many studies have introduced metrics focusing on analysing passenger waiting and travel times.

2.2.2. Passenger waiting and travel time

Before the emergence of new automatic data collection technologies such as AFC and APC, traditional frameworks, with several methods to analyse passenger waiting time, have been developed based on basic data, e.g., AVL data from traffic control systems (NASEM, 2006). Two classes of metrics have been developed in the literature using such data, namely: timetable-based measures of deviations, mainly used for low/medium-frequency services, e.g., OTP and CPM; and headway-based measures mainly used for high-frequency services (Zhao et al., 2013). These methods allow for studying how schedule and headway deviations (unpunctuality) and their uncertainties affect both passengers' actual and budgeted waiting and travel times. For instance, the budgeted waiting time is often divided into *ideal* waiting time, which would result from service exactly following the schedule; and *excess* waiting time which is the difference between *actual* and ideal waiting time and is due to service unpunctuality (NASEM, 2006).

In their study of optimal running time schedules, Furth and Muller (2007) mention two metrics directly impacting service reliability and punctuality, i.e., *excess waiting time* (EWT) and *potential travel time* (PTT). EWT reflects the excess waiting time that passengers incur due to unpunctual departure times whereas PTT, also called buffer time, relates to the budgeted travel time but not used in waiting or riding. **Figure 4** illustrates EWT (wait) and PPT (potential) times, and typical distributions for a passenger service departing from i and arriving at j.

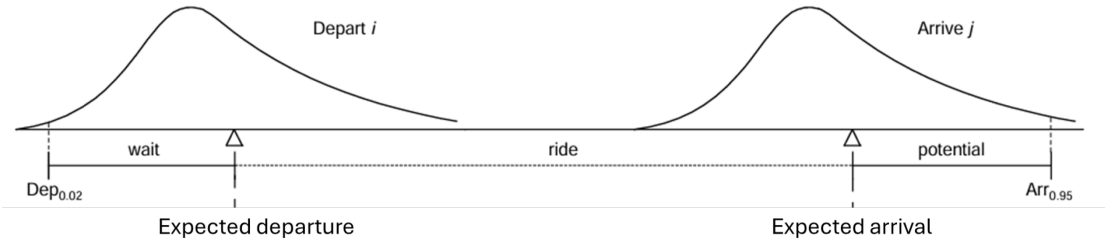


Figure 4. Illustration of a typical departure and arrival distribution of vehicles for analysing passenger waiting and travel times, adapted from (Furth and Muller, 2007).

Focusing on analysing passenger waiting times, basic AVL data help analyse platform waiting time $W_{platform}$, i.e., time passengers spend waiting at a stop/station platform, and budgeted waiting time, which consists of $W_{platform}$ and potential waiting time $W_{potential}$ depending on the risk passengers are willing to take for missing a departure (NASEM, 2006).

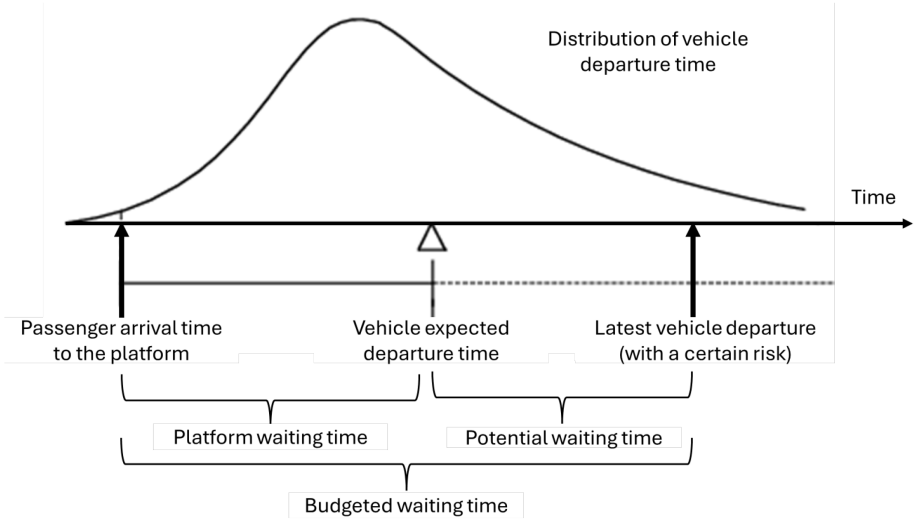


Figure 5. Illustration of a framework for passenger waiting time analysis.

These are often combined into one equivalent metric called the *equivalent waiting time* $W_{equivalent}$ (NASEM, 2006). In equation (2), $W_{equivalent}$ is the average between $W_{platform}$ and the 95th-percentile waiting time (often noted $W_{.95}$). In this formulation, passengers are assumed to accept a 5% risk of missing their departure.

$$W_{equivalent} = \frac{1}{2}(Wplatform + .95) \quad (2)$$

To better analyse the waiting and travel time of passengers, it is important to consider and account for how they arrive at the departure stations/stops. Passengers adjust their arrival at the station/stop based on, among others, knowledge of the schedule, punctuality and reliability performances (Wilson et al., 1992). Therefore, existing studies generally distinguish between the analysis of long- and short-headway services. In the former, passengers are expected to arrive at stops just before the scheduled departure time to reduce their waiting time, whereas, in the latter, they are usually assumed to arrive randomly without looking at the schedule, i.e., even if there is a one (Furth and Muller, 2007).

2.2.3. Short-headway services

Short-headway routes and services are found in many PT systems including urban commuter rails and many bus systems in the inner portions of major cities, particularly in peak periods. Assuming that passengers' arrival rate is evenly distributed (i.e., constant arrival rate) and that they board the first vehicle that arrives, Wilson et al. (1992) show that the *expected waiting time* EW is given by equation (3), where $E[H]$ and $CoV[H]$ are mean and coefficient of variation of actual headway, respectively. The latter is calculated as the standard deviation of headway divided by the mean $E[H]$.

$$EW[H] = \frac{1}{2}E[H](1 + CoV[H]^2) \quad (3)$$

In the case of perfect headway adherence, EW for passengers is half the headway, but it increases as the headway variation increases. Based on basic AVL data and in the absence

of more detailed passenger data, equation (3) estimates passengers' waiting time assuming a constant rate of passenger arrivals. To reduce errors from this assumption, this measure can be studied for short periods, e.g., peak hours, and should therefore be avoided for analysing longer periods, e.g., a whole day (Wilson et al., 1992).

A related measure that is a useful supplement to EW is EWT which was also proposed by Wilson et al. (1992). The authors define EWT as the difference between the actual passenger waiting time and the expected waiting time that would result from perfect adherence to schedule, see the formulation in equation (4). EWT is especially useful when comparing service quality across routes with quite different headways.

$$EWT = EW[H] - EW[H_{scheduled}] \tag{4}$$

When focusing on analysing specific periods, e.g., short headways or peak hours, where $H_{scheduled}$ is constant, EWT is simply $EW[H]$ minus half of $H_{scheduled}$. In most cases, e.g., longer periods, $H_{scheduled}$ will be variable. Used for monitoring PT reliability by London Transport, EWT was extended later to the entirety of journeys (comparing mean actual and schedule values of each component of passenger journeys) to capture service reliability and compare it across routes with quite different headways (Zhao et al., 2013).

Even in the absence of data on passenger arrival rates, it is possible, at least for high-frequency services, to measure how well passengers are served at their departure stops, Assuming random passenger arrivals (valid for short-headway services), Wilson et al. (1992) define the percentage of passengers receiving good service based on how much of the waiting time for the next vehicle is within $H_{scheduled}$. For bad services, they consider how much of the excess waiting time (to the next vehicle) is more than $2 \times H_{scheduled}$.

2.2.4. Long-headway services

In the case of services with long headways, passengers are assumed to arrive at stops at times designed on the timetable to minimize their waiting time, i.e., just before the scheduled departure time. Hence, schedule deviation $\delta D := D - D_{scheduled}$ are the main determinant of the passengers' waiting time, where $D_{scheduled}$ and D are the scheduled and actual departure times, respectively. Thus, negative values of δD represent earlier vehicle departures.

As mentioned earlier, Furth and Muller (2007) introduced EWT and PTT, see **Figure 4**. The authors suggest, in the case of long headways, using 2nd-percentile departure times $D^{(.02)}$ to measure EWT, and 95th-percentile arrival times $A^{(.95)}$ for PTT. For a stop s, EWT_s and PTT_s are reformulated in equation (5) and (6), respectively.

$$EWT_s = E[D_s] - D_s^{(.02)} \tag{5}$$

$$PTT_s = A^{(.95)} - E[A_s] \tag{6}$$

In both equations, $E[X]$ refers to the expected value of the random variable D_s and A_s representing the actual departure and arrival time at a station/stop s, respectively. In these formulations, passengers are assumed to accept a 2% risk of missing their departure, and a 5% risk of arriving late to their destination.

Focusing on analysing waiting times, the short-headway framework is extended to study the effect of schedule deviations on passengers' waiting time for long-headway services.

As illustrated in **Figure 6**, the excess budgeted waiting time is the spread in the schedule deviation distribution, i.e., $\delta D^{(.95)}$-$\delta D^{(.02)}$.

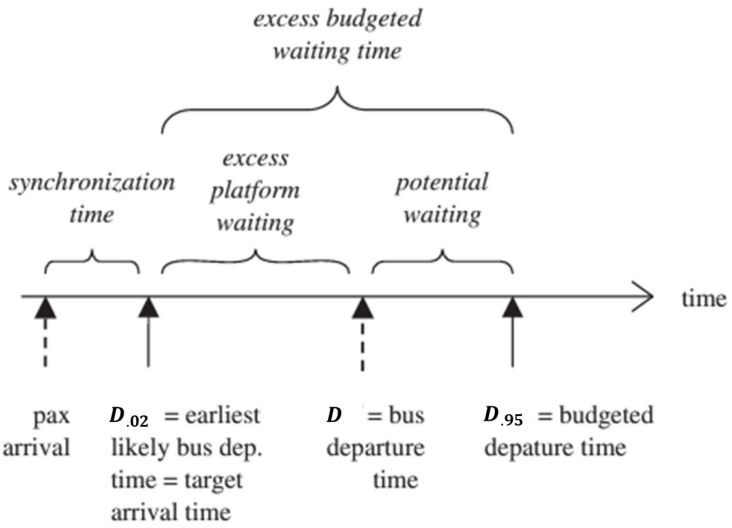


Figure 6. Illustration of the framework for analysing passenger waiting time for long-headway bus services, adapted from (NASEM, 2006).

The excess budgeted waiting time, in **Figure 6**, includes both excess platform and potential waiting time which relates to EWT and PTT, respectively. These can be combined in an equivalent excess waiting time, i.e., $W_{equivalent}$ using the previous formulation in equation (2). The synchronisation time is independent of schedule deviations as it is an inevitable consequence of, e.g., human risk aversion, headway service design, and uncertainty in station access times (NASEM, 2006).

2.2.5. Other metrics

More generally, other alternative metrics have been commonly used to assess punctuality, e.g., using continuous metrics that reflect actual delays, such as total or average delays, variability metrics (variance, standard deviation and CoV), statistical distributions of running time and schedule deviations (Zhao et al., 2013). Other metrics have been studied and introduced such as the risk of delay that is proposed by Ferreira and Higgins (1996) to capture the probabilities of specific delay durations, and other retrospective metrics such as maximum delay and delay-at-risk (Blayac and Stéphan, 2021).

Common metrics such as total or average (vehicle) delays, variability metrics, and distributions, provide valuable insights from an operational perspective but may not fully capture passengers' experiences and expectations. For instance, Börjesson and Eliasson (2011), and more recently Denti and Burroni (2023), criticized the use of "average delay" as a measure of (train) reliability as it does not reflect passengers' preferences. Moreover, service quality frameworks, such as EN-13816 by CEN (2002), underscore the importance of incorporating user-oriented perspectives into punctuality metrics. While operator-centric metrics like OTP remain prevalent, there is a growing recognition of the need for passenger-centric metrics that directly measure punctuality from the perspective

of passengers. This shift towards passenger-focused metrics aims to align service quality assessments more closely with passenger patterns and expectations, ultimately enhancing the overall passenger experience and satisfaction (Danaher et al., 2020).

2.3. Metrics for passenger punctuality

The punctuality metrics discussed thus far have primarily focused on vehicle-based assessments rather than directly reflecting the experiences of passengers. Many studies emphasize the importance of adopting a user-centric perspective when measuring punctuality in PT systems (Wardman, 2001, Mishalani et al., 2006). There is therefore a notable shift in recent years towards using passenger-oriented measures, particularly in quantifying passenger punctuality (Hendren et al., 2015).

Monitoring passenger punctuality involves assessing how well passengers reach their destinations within predefined time windows around scheduled times, encompassing their entire journey from origin to destination, including transfers (Parbo et al., 2016). For instance, Lee et al. (2014) identify, at a particular transfer, that scheduled transfer time, distributions of vehicle arrivals, headways and walking time have major effects on passenger punctuality. Hence, unlike vehicle punctuality, which evaluates trips of individual vehicles, passenger punctuality emphasizes the holistic experience of passengers throughout their travel itinerary. Defining and characterizing passenger punctuality is therefore less straightforward compared to vehicle punctuality metrics (Nelldal et al., 2019).

One straightforward way to assess passenger punctuality of PT services is to conduct manual customer surveys where passengers can give feedback and are directly asked to rate the punctuality of the PT services they use. Initially, such surveys enabled transit agencies to directly observe a limited part of the passenger experience due to its high cost of collection and processing, limiting sampling frequency and system coverage. This changed with the arrival of technologies for automatic data collection, such as AFC and APC, creating large sets of data on individual vehicle movements and passenger demand that could be used to estimate the passenger experience more efficiently (Uniman et al., 2010).

In the following, we will explore specific metrics that can potentially be used to monitor passenger punctuality in PT systems. Unlike the previously reviewed vehicle-centric measures, these metrics can better reflect passengers' experiences and perceptions, and may therefore provide more valuable insights to improve the reliability of PT systems.

2.3.1. General demand estimates

While OTP is commonly used as an operator-oriented metric, it can be refined to be more user-oriented by incorporating weights that represent passenger volumes. For instance, Kristoffersson and Pyddoke (2019) studied punctuality from a train passenger perspective using ridership data for some Swedish regional lines. The authors compared train punctuality, measured as RT+5, with passenger punctuality, measured as the share of passengers arriving at their destination within 5 minutes. The authors distinguish between punctual (RT+5) and significantly late/unpunctual (RT+30) train/travellers. Similarly,

Nelldal et al. (2019) measured passenger punctuality on a larger Swedish rail network (different types of trains, lines and time periods) by weighting the percentage of delayed trains by the average number of passengers. The authors explored several related measures, e.g., total and average passenger delay (in pax-min), and noted that the breakdown by the hour, rather than by individual trains, allows for calculating punctuality more effectively with existing data. However, accurately determining the number of passengers per individual train is challenging. Despite this, estimating the distribution of passengers per hour across different day types is feasible, albeit not precise for each train. The authors reported total and average delay time experienced by passengers on delayed trains excluding cancelled trains.

Unlike Kristoffersson and Pyddoke (2019) who found no significant difference between passenger and train punctuality results, Nelldal et al. (2019) highlight their differences which are more pronounced during peak hours and on longer routes. As mentioned earlier, discrete metrics like OTP or CPM, although made more user-oriented, have limitations in capturing the magnitude of delays and therefore nuances of passenger experience (Barabino et al., 2015).

To monitor passenger punctuality and measure the service performance of different train contract groups in British railways, NetworkRail (2017) uses composite metrics such as *Average Passenger Lateness* (APL) and *Total Passenger Lateness* (TPL). These are an estimate of how late every passenger reaches their destination station. APL is calculated by determining the lateness of trains at monitoring points relative to scheduled times, incorporating cancellation factors and passenger weights, and then averaging these values per day and service group. TPL aggregates the lateness values over a period for each service group, reflecting the overall cumulative lateness experienced by passengers across multiple days. The expected number of passenger journeys, i.e., demand forecast, is calculated annually based on ticket sales and revenue data, and adjustments are made within the year.

Similarly and based on national demand prognoses, the Dutch IM (*ProRail*) and national operator (NS) have monitored passenger punctuality in the realized train timetable (Wolters, 2016). The PP_1 metric was used until 2015 and is formulated in equation (7), where $A_{realized}$ and $T_{realized}$ are, respectively, the arrivals and transfers that are successfully realized.

$$PP_1(\tau) = \frac{A_{realized}(\tau)+T_{realized}(\tau)}{A_{planned}+T_{planned}} \tag{7}$$

The denominator accounts for the corresponding planned quantities according to the scheduled timetable. A certain arrival is realized if it is within the delay threshold τ from the planned arrival. Transfers are only considered at stations with high numbers of (forecasted) transferring passengers and are realized if transfer time is at most 7 minutes and at least enough to cross the platforms. As mentioned by Wolters (2016), these quantities are measured at around 35 stations and weighted with demand estimates, alternatively, an average train ridership is assumed. The demand forecasts are typically made for different periods, i.e., morning/evening peak hours and off-peak hours on workdays, and weekends.

2.3.2. Detailed passenger trip data

As mentioned earlier, automatic data collection systems, e.g., AFC, and APC, offer valuable insights into passenger demand and behaviour, and hence passenger punctuality. These technologies are increasingly used in PT systems to capture detailed passenger journey information (Uniman et al., 2010). Figure 7 illustrates how detailed AFC data can be matched to the passenger journey time. This enables, therefore, more accurate estimations of travel experiences of individual origin-to-destination (OD) passengers, thus providing a new, easy-to-access resource for the monitoring, evaluation, and analysis of service quality and enhancing the understanding of passenger experiences including punctuality.

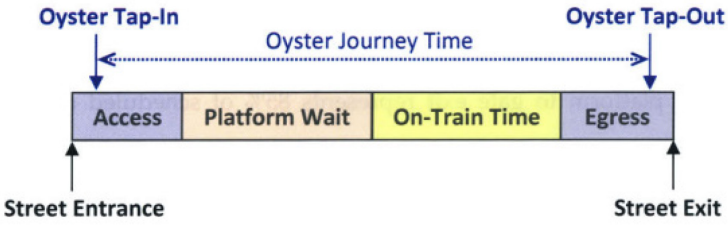


Figure 7. Illustration by Chan (2007) of the components of the passenger journey time and use of tap-in/out information from (London's Oyster) smart card data.

By combining, for instance, AFC data from smart card transactions and AVL data tracking vehicle movements, it is possible to not only improve the accuracy of existing measures, e.g., the one based on general demand estimates, but also allow for the definition of new metrics that better represent punctuality for PT users (Bagherian et al., 2016). Pelletier et al. (2011) emphasize the diverse applications of data from AFC systems, such as smart cards, in the management of PT systems, particularly in calculating performance and reliability indicators at the tactical level, especially in systems with entry and exit validations (Uniman et al., 2010).

In their early study of passenger punctuality of high-frequency PT services, Wilson et al. (1992) considered accounting for variable passenger arrival rates. The authors suggested an improved measurement of EW and previously formulated in equation (3), by incorporating passenger weights in the formulation. Let n be the number of consecutive vehicle trips from a certain stop, and H_i is the headway of the i^{th} vehicle departure. By noting p_i as the mean arrival rate (e.g., in pax/min) of passengers to the station/stop before the i^{th} departure, the *expected (passenger-weighted) waiting time* EW_p can be formulated as in equation (8).

$$EW_p(H) = \frac{1}{2}\frac{\sum_{i=1}^{n} p_i H_i^2}{\sum_{i=1}^{n} p_i H_i}, \tag{8}$$

In theory, EW_p can better measure passenger waiting time including over longer periods with highly variable passenger arrivals (Wilson et al., 1992). However, this measure requires access to more detailed data on passenger arrival rates. This is the case, for instance, using AFC data at PT train stations, unlike bus passengers who only use their tickets when boarding the vehicle (not when arriving at the bus stop).

Using AFC data, i.e., time-stamped Oyster smartcard transactional data from the London rail network, Chan (2007) introduced and applied a new metric named *Excess Journey Time* (EJT) to measure actual journey times experienced by passengers. For a given OD pair od, EJT_{od} is defined using the difference between actual passenger journey times A_k and scheduled journey time S_{od}, or some other pre-defined journey time standard, see the illustration in Figure 8 by Chan (2007).

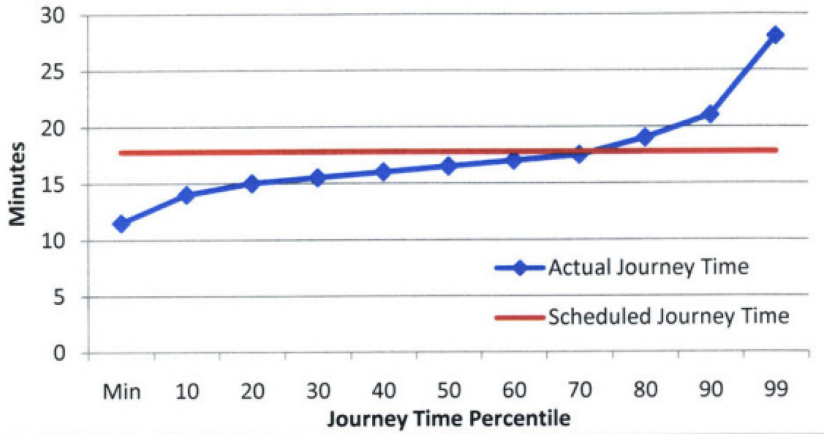


Figure 8. Typical distribution of journey times for an OD pair (Chan, 2007).

Given $|k|$ passenger trips on an od pair, equation (9) formulates the EJT_{od} score. This metric can be extended to a complete line by weighting the scores from individual pairs with their passenger ridership. EJT is hence a measure that balances the passenger's and operator's perspectives of PT service quality.

$$EJT_{od} = \frac{\sum_{k \in od} \max(A_k - S_{od}, 0)}{|k|}, \tag{9}$$

The same author also introduced the *Journey Time Reliability* (JTR) metric to quantify service reliability as experienced by passengers based on the observed journey time distributions. EJT has, however, found more applications such as Zhao et al. (2013) who applied at various levels of spatial and temporal aggregation to measure and evaluate the service quality for different London Overground lines. The authors' results show substantial variations across the different lines and times of weekday service.

Using detailed AFC data (from the Oyster smartcard ticketing system) to measure the service reliability of the London Underground, Uniman et al. (2010) introduce the *reliability buffer time* (RBT) measure as the extra time that passengers need to add to their usual travel time to make sure they reach their destination with a certain probability. RBT is formulated as the difference between the 95th and 50th (median) percentiles of the travel time. This metric is aggregated from the OD pair to the line or network level using the OD flow-weighted average. The author developed another variant called *Excess Reliability Buffer Time* (ERBT) to distinguish between typical and non-recurring delays, and hence between incidents or disruptions. Uniman et al. (2010) define ERBT as the extra buffer time passengers need to arrive on time with 95% certainty, besides the normal buffer time for typical conditions $\text{RBT}_{\text{typical}}$. **Figure 9** provides an illustration of the ERBT as the difference between $\text{RBT}_{\text{overall}}$ and $\text{RBT}_{\text{typical}}$. As presented in the figure, it

is also possible to estimate the proportion of trips that are unpunctual given the accepted reliability standard (95 %). These measures are, however, more appropriate in the context of high-frequency services where passenger arrival pattern is considered uniform (Bagherian et al., 2016).

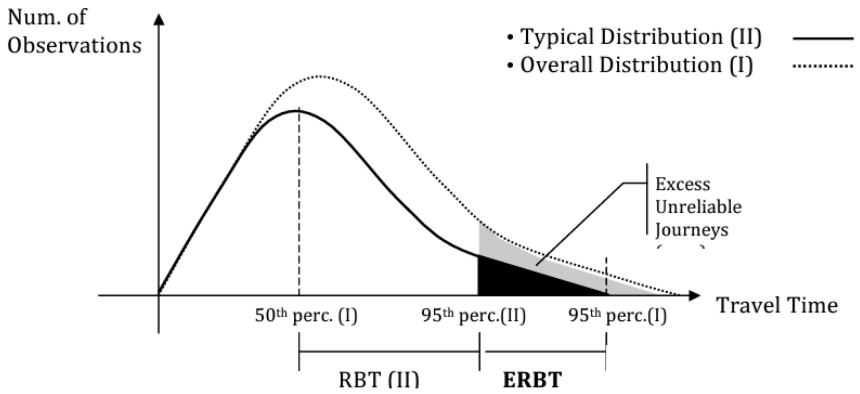


Figure 9. Illustration of the RBT and ERBT measures by Uniman et al. (2010).

Given the availability of better passenger demand data from AFC systems, NS has used a new method to calculate passenger punctuality since 2015 (Wolters, 2016). The new method addresses the limitations of the previous metric, in equation (7), that is based on forecasted demand, e.g., limited station and transfer coverage and peak hour disparities. Based on check-in times from PT smart cards, the new metric is simply the ratio of promised and total trips, see the formulation in equation (10), where T_{total} is the total number of trips, and $T_{promised}(\tau)$ is the number of trips within the promised total travel time plus a threshold τ.

$$PP_2(\tau) = \frac{T_{promised}(\tau)}{T_{total}} \tag{10}$$

$T_{promised}$ can be directly inferred from the check-in time of every trip which is used to determine the promised journey for the individual OD passenger. The promised journey is linked to the earliest possible realized arrival time if the next, after check-in time, scheduled departure is taken (Wolters, 2016).

In their research on passenger-oriented measures of reliability, Bagherian et al. (2016) introduce two measures, i.e., one focusing on punctuality (deviation from the schedule) and another on predictability (day-to-day variation). The former, which is more relevant here, measures the *schedule deviation* (SD) of individuals' actual travel time from the scheduled one. Given an OD pair and based on a sufficient sample of AFC data on this pair, SD is defined as the ratio of the excess delay to travel time at a targeted percentile σ, e.g. median ($\sigma = .5$) or 95$^{\text{th}}$-percentile ($\sigma = .95$).

Let (i,j) be a given OD pair, the scheduled deviation measure $SD^{(i,j)}$ can be formulated as in equation (11), where $T^{(i,j)}$ is the actual/observed distribution of passengers' travel times between OD pairs (i,j). $\delta T^{(i,j)}$ is the distribution of the deviation between the actual and scheduled passenger travel times, i.e., excess delay distribution.

$$SD^{(i,j)} = E[\frac{\delta T_{\sigma}^{(i,j)}}{T_{\sigma}^{(i,j)}}] \quad (11)$$

Since passenger trips must be matched to the service schedule/timetable, the calculation of the measure often requires more parameters and computations, e.g., trips with transfers, transfer points, alternative routes, and walking times. Bagherian et al. (2016) applied the measure to the public transport network of The Hague, the Netherlands. The authors computed the SD measures for specific values of σ and at various spatial-temporal levels, from a single OD pair to a network-wide evaluation using weighted averages.

2.3.3. Socio-economic assessment

Another possible approach for monitoring passenger punctuality is using the economic costs of passengers' unpunctuality. In their comprehensive assessment of the costs of flight delays in the United States, Ball et al. (2010) account for different cost components such as passenger costs comprising delays, cancellations, missed connections, and other costs for the operators. Similarly, the reliability of PT services can be assessed using economic costs, e.g., by multiplying the minutes of service delay with the cost per minute delay., the costs per minute of delay comprise both the direct costs to service providers per minute and the economic cost to late passengers per minute.

In their study of how PT service reliability can be incorporated into economic assessments, such as *cost-benefit analysis* (CBA), van Oort (2016) demonstrates how to calculate the impact on passengers. The author combined variants of previously presented metrics, e.g., $EW[H]$, with existing economic valuation parameters such as the value of time (in Euros/hour) and the value of reliability (in Euros/hour of standard deviation). These valuation parameters are generally different for different trip purposes, e.g., business, commuting or other trips. Based on simulated AVL and APC data, van Oort (2016) illustrates the approach in a CBA application on a PT line in Utrecht, The Netherlands.

In an extensive review of passenger-centric rail planning, a related common approach from the literature is mentioned by Parbo et al. (2016) which emphasizes travel behaviour. Quantifiable attributes, such as valuation of travel and waiting time, headway, and delays, affect passengers' travel behaviour. These are often modelled as components of a generalized travel cost function (Parbo et al., 2016). Using extensive data and/or simulations allows for more accurate monitoring of these valuations, and hence the impact of unpunctuality on passengers' travel behaviour and perception of PT service reliability.

2.3.4. Bus services

Methods for monitoring passenger punctuality in rail transport often involve comparing actual arrival times with scheduled times at specific stations. In contrast, bus services present unique challenges, such as bus overtaking, which can disrupt the scheduled order of arrivals. Barabino et al. (2015) propose a passenger punctuality measure based on AVL data, focusing on the first bus to arrive at a stop, even after overtaking. The measure is defined as the fraction of passengers who will be served within an acceptably short interval after they arrive. This method provides a more accurate assessment of passenger

punctuality and enables monitoring in each direction of a bus route in a transit network, as well as for every bus stop and time interval.

Focusing on bus transit, Diab et al. (2015) reviewed the differences between passengers' and transit agencies' perspectives on bus service reliability including punctuality. The authors identified the main gaps between these two perspectives for planners and decision-makers aiming for reliability improvement strategies for bus services. An important gap is the drawbacks of OTP standards in capturing variations in waiting and running times which affect more travel decisions of bus riders.

By reviewing (more than twenty) measures of bus service reliability from the travellers' perspective, Gittens and Shalaby (2015) found that they are generally based on monitoring arrivals at the destination, waiting times at the origin stop, and consistency in waiting and travel times. However, the authors highlight that no measure captures all these aspects. A new composite indicator, i.e., *journey time buffer index* (JTBI), is introduced combining aspects of travel time (such as RBT), schedule adherence (such as OTP), headway regularity (such as CoV), and waiting time (such as EW). The authors formulate JTBI by distinguishing between long- and short-headway bus services. As a composite metric, the JBTI formulations include an arrival penalty (accounting for bus travel time variability) and a wait penalty (capturing the variability in departure times at the origin stop). Although lacking detailed AFC and APC data, Gittens and Shalaby (2015) show that JTBI is better suited, compared to the reviewed metrics, to identify the factors contributing to unreliable bus service from a passenger perspective.

2.4. Summary and discussion

As passenger punctuality in PT systems affects travellers' experience and operators' operational efficiency, a number of alternative metrics have been developed to monitor the punctuality and reliability of PT services from passengers' perspectives. For passengers, punctuality impacts their travel experience, including waiting and travel times, and comfort. For operators, it is linked to operational efficiency and economic costs. OTP is commonly used by PT operators and agencies but has limitations, e.g., not capturing the severity of delays or passenger punctuality. Various metrics, other than OTP, e.g., DI, CPM and ESA, have been introduced and employed to assess PT punctuality. However, better monitoring and improving passenger punctuality can enhance both satisfaction from a passenger perspective and PT ridership and long-term economic performance from an operator's point of view.

Choosing appropriate metrics for monitoring passenger punctuality in PT systems is essential. Frameworks like EN-13816 highlight the importance of incorporating passenger perspectives in monitoring PT service reliability. Therefore, more passenger-oriented metrics, like EWT and PTT, have been first developed to offer better alternatives for monitoring punctuality. These metrics aim to capture different aspects of punctuality and service reliability impacting passengers' travel experience, e.g., passengers' waiting and travel times.

Thanks to the emergence of new data collection systems, there has been a shift towards developing more passenger-oriented metrics to assess how well passengers reach their

destinations within predefined time windows, considering their entire journey. Methods based on general demand forecasts and more detailed automatic data collection (AFC, AVL, APC) provide valuable improvements in monitoring passenger experiences including punctuality. Metrics such as APL, RBT, and EJT are used to evaluate service quality from a passenger perspective. Economic-based methods, and others specific to bus traffic, can further improve measurements of passenger punctuality.

3. Qualitative comparisons

One of the main objectives of this study is to explore different measures of passenger punctuality that can be used to capture the reliability performance of PT systems. In the previous section (section 2), we reviewed some existing literature and practices on measuring punctuality in different transport systems and identified several punctuality metrics, including passenger-centric ones, that could be applicable for our purpose, see **Table 2** for an overview.

Table 2. Overview of reviewed metrics for monitoring punctuality in PT systems.

Categories	Passenger perspective	Metrics
OTP-variants	No	On-Time Performance (OTP), Time-to-X, En-route Schedule Adherence (ESA), Delay Increment (DI), Combined Performance Measure (CPM).
	Yes	Passenger-weighted OTP, passenger punctuality scores (PP_1 and PP_2), Average Passenger Lateness (APL), Total Passenger Lateness (TPL)
Passenger waiting & travel time	No	Equivalent waiting time $W_{equivalent}$. **Short headways:** Expected waiting (EW), Excess Waiting Time (EWT), reliability buffer time (RBT), Excess Reliability Buffer Time (ERBT). **Long headways:** Potential Travel Time (PTT)
	Yes	**Short headways:** Expected (passenger-weighted) waiting time (EW_p), OD-aggregated RBT and ERBT. **Long headways:** Excess journey time (EJT), schedule deviation (SD)
Miscellaneous	No	**Other tools:** Total/average delay, Distributions (running time, schedule deviations), Risk of delay, Standard deviation, Coefficient of variation, Maximum delay, Delay-at-risk.
	Yes	Total or average passenger delays. **Economics** (using passenger demand): Value of reliability, generalized travel costs. **Bus-specific** (using passenger demand): Fraction of served bus passengers served (within a threshold), journey time buffer index (JTBI).

In this section, we will conduct a more detailed qualitative assessment of these metrics, based on different aspects for use in PT systems such as relevance, effectiveness, challenges, and potential. We will also compare them with each other and with the conventional metrics such as OTP, which is widely used by PT operators and authorities. Considering data availability for the case study, this assessment ends by selecting some suitable measures for further quantitative analyses in section 4.

3.1. Relevance for PT services

One important aspect in assessing passenger punctuality metrics is its relevance to PT systems in general. Given the specificities of PT systems compared to other transport systems, relevant passenger-centric metrics should be able to capture the punctuality in the passenger travel experience throughout the entire journey. For instance, most of the reviewed passenger-centric metrics differ from the traditional OTP, which only measures the schedule adherence of the PT vehicles often at terminal stations/stops, without

considering relevant aspects for PT systems such as the passengers' travel times, waiting times, transfers, crowding, and comfort. While OTP may be relevant for the planners of the PT operators and agencies, it is less relevant for reflecting the PT passenger experience, which may vary depending on certain characteristics of the studied PT system, e.g., service frequency, passenger demand, and network coverage.

Although applied to monitor punctuality on passenger rail transport, the reviewed OTP variants, such as CPM, DI/DC, and Time-to-X, are not passenger-centric since they capture schedule deviations of trains rather than passengers. While such metrics can be used for rail infrastructure management, e.g., monitoring infrastructure failures and performance regimes, they are, as such, less relevant for use in PT systems to monitor passenger punctuality and service reliability. Although also an OTP variant, the ESA metric was adopted for monitoring the punctuality of the PT system in New York, but was replaced soon after as it is less suitable for short-headway PT services (Cramer et al., 2009). Based on demand estimates or forecasts, some OTP-variants, e.g., PP_1 and passenger-weighted OTP, are used to improve the relevance of the metric for monitoring punctuality in passenger transport including PT systems.

The majority of the reviewed metrics are passenger-centric or intended proxies for it since they can (in-)directly capture some aspects of the passenger journey experience, such as passenger waiting and travel time. This passenger-centricity is an important aspect of the relevance of the reviewed metric. **Table 3** presents an overview of relevant metrics for passenger-centric monitoring of punctuality in PT systems. For each metric, the table describes the corresponding practical/theoretical application and passenger transport system.

Table 3. Comparison of the relevance of certain metrics for monitoring punctuality in PT systems.

	Use in practice or research	Transport system	Unit
APL & TPL	Applied by the British IM, e.g., for performance metric/regimes of different train/contract groups and periods.	National passenger rails	Time
$\boldsymbol{PP_1}$ & $\boldsymbol{PP_2}$	Applied by Dutch IM for monitoring train passenger punctuality.	National passenger rails	Percentage (%)
$\boldsymbol{EW}$	Theoretically proven for randomly arriving passengers at a constant rate.	PT (high-frequency)	Time
$\boldsymbol{EW_p}$	Introduced in the research as a better alternative to EW given accurate data on passenger arrival rates	PT (high-frequency)	Time
$\boldsymbol{EWT}$	Applied by London Transport to monitor service reliability.	PT	Time
$\boldsymbol{EWT_s}$ & $\boldsymbol{PTT_s}$	Theoretically defined 2-percentile departure times and 95-percentile arrival times, respectively.	PT (low-frequency)	Time
$\boldsymbol{RBT}$ & $\boldsymbol{ERBT}$	Introduced in the research and illustrated in a case study from the London Underground network.	PT (high-frequency, metro)	Time
$\boldsymbol{EJT}$	Introduced in the research and illustrated in a case study from the London Overground network.	PT (rail)	Time
$\boldsymbol{JTBI}$	Introduced in the research and illustrated in a case study from a bus network in Ontario.	PT (bus)	Normalized index (%)
$\boldsymbol{SD}$	Introduced in the research and illustrated a case study from the PT network in the Hague.	PT	None (ratio)

A few of the metrics in **Table 3** have already been applied and used in practice. Apart from EWT and metrics with applications in the passenger rail systems, e.g., for monitoring train passenger punctuality and train contract performance, all of the reviewed

metrics in **Table 3** are so far only introduced in the research and illustrated in a (real-world) case study for different type of PT systems.

The relevance of most of the introduced metrics is often related to the headway/frequency of the PT system. For instance, EWT_s and PTT_s are introduced to be applied in the case of low-frequency PT services, whereas RBT and $ERBT$ are more relevant when monitoring services with short headways such as during peak hours. Other metrics have been constructed to be applied to specific PT systems, e.g., $JTBI$ for bus traffic. Note, however, that a metric might be relevant for use in certain PT services even if there are no reviewed applications or case studies. For instance, although applied to intercity passenger rail, PP_1 and PP_2 can be relevant for use for PT systems such as local commuter rails.

When assessing the relevance of these metrics for use in PT systems, it is important to consider the measurement unit (last column in **Table 3**) and whether it allows for cross-service and cross-mode comparisons. PT-relevant metrics should be scalable across different levels, e.g., from stop, route & trip to all systems, and consistently enable comparisons among routes/lines and between different PT modes and systems.

3.2. Monitoring punctuality of PT passenger

To effectively and accurately monitor passenger punctuality, metrics should represent the real impact that is ultimately experienced by the passengers, and therefore go beyond assessing the PT service provision. In other words, these metrics should objectively capture the real passenger experiences and reflect the service punctuality from the passenger's viewpoint. Effective metrics should therefore comprehensively capture all (or the main) aspects of the passenger journey experience. Moreover, they should also be useful for the PT agency and other stakeholders in that they provide actionable insights into how the service quality can be improved, see the next subsection for more on this. Accurate metrics should, however, objectively and precisely capture the details of the passenger journey experience including waiting times, travel times, and transfers.

To assess the effectiveness and accuracy of the reviewed metrics, it is important to compare them in terms of the aspects that each metric is capturing from the passenger journey experience, the most disaggregated level of the metric, and whether it is based on an improvement of one or more other metrics. Moreover, schedule adherence is not always a major concern from a passenger's point of view as passengers perceive headway adherence or service regularity as more important in high-frequency PT services such as during peak hours (Parbo et al., 2016). Focusing on these elements, **Table 4** compares the effectiveness and accuracy of reviewed passenger-centric metrics in monitoring passenger punctuality in PT systems.

Several metrics have been developed to provide a more passenger-centric perspective of punctuality compared to traditional metrics such as OTP. As indicated in **Table 4**, metrics like PP_1 aim to improve OTP by focusing on journey times including transfers which are more critical to the passenger experience than vehicle delay at certain intermediate or

final stops. Moreover, PP_1 operates at a higher level focusing on specific routes with transfers to/from other routes. Given more detailed AFC data, the latter can be further improved, using PP_2, to operate at the passenger trip level to capture more accurate individual journey times.

Table 4. Effectiveness and accuracy of reviewed passenger-centric metrics in monitoring passenger punctuality in PT systems.

Metric	Captured from passenger travel experience	Minimal level	Based on (if any)
$\boldsymbol{PP_1}$	Journey time	Route	$\boldsymbol{OTP}$
$\boldsymbol{PP_2}$	Journey time	Trip	$\boldsymbol{PP_1}$
$\boldsymbol{EW}$	Waiting time (peak hours)	Stop (departure)	
$\boldsymbol{EWT}$	Waiting time	Stop (departure)	$\boldsymbol{EW}$
$\boldsymbol{EW_p}$	Waiting time (peak hours)	Stop (departure)	$\boldsymbol{EW}$
$\boldsymbol{EWT_s}$	Waiting time (off-peak hours)	Stop (departure)	$\boldsymbol{EWT}$
$\boldsymbol{PTT_s}$	Travel time (off-peak hours)	Stop (arrival)	
$\boldsymbol{RBT}$	Travel time (typical)	Trip	$\boldsymbol{PTT_s}$
$\boldsymbol{ERBT}$	Travel time (atypical)	Trip	$\boldsymbol{RBT}$
$\boldsymbol{SD}$	Travel time	Route & trip	$\boldsymbol{PTT_s}$
$\boldsymbol{EJT}$	Journey time	Trip	
$\boldsymbol{JTBI}$	Peak or off-peak journey time	Trip	$\boldsymbol{EW, RBT}$

Research has shown that waiting time is perceived by PT passengers to be more important than in-vehicle time, especially during peak hours (Wardman, 2001, Mishalani et al., 2006). It is therefore important to accurately capture such aspects of the passenger journey using metrics such as EW, EW_p, and EWT which are particularly suitable when studying specific PT travel periods, e.g., peak or off-peak hours. EW captures peak-hour waiting times at passengers' departures. Given accurate data on passengers' arrival rates, while EW_p improves the latter by operating at different travel periods. In the absence of such data, EWT is an improvement of EW as it also operates during off-peak travel periods. EWT_s is suggested as an improvement of EWT but is more suitable for long-headway PT services during off-peak hours. In this case, PTT_s is a complementary metric to capture the passengers' budgeted travel times, i.e., part of the passenger journey other than waiting times.

To gain more insights into the reliability of passengers' travel times, metrics like RBT and $ERBT$ are suitable at the trip level. Unlike the former which focuses on typical or recurrent delays, the latter is more accurate for capturing non-recurrent scenarios. Both metrics can be used to understand how much additional time passengers budget for potential delays. Operating at both route and trip levels, SD focuses on the overall travel time variations and their impact on passenger experiences. Including all of the passenger journey time, EJT provides a more comprehensive assessment compared to the previous metrics focusing on either waiting or travel times.

Building upon the previous metrics, $JTBI$ combines elements from EW and RBT, focusing on journey times either during peak or off-peak hours. Although developed for bus services, $JTBI$ allows therefore to capture at the same time passengers' waiting and travel time variations at both route and trip levels.

3.3. Implementation in PT systems

When assessing the reviewed metrics, it is also essential to consider their usability and practicality. Ease of implementation is a key usability aspect which involves, among others, the type and amount of data required, calculation complexity, and integration with systems already in place. Another factor is the operational feasibility encompassing support for real-time monitoring, scalability across different PT systems, as well as maintenance. Additionally, interpretability and actionability are important, ensuring that metrics provide clear, actionable insights and can be effectively communicated to all PT stakeholders. Cost-effectiveness and adaptability are also essential, weighing the costs of implementation and operation against the possible benefits, and making sure the metrics can be adjusted to different contexts. See **Table 5** for a comparative overview of some of these main aspects with a focus on the improved versions of the reviewed metrics.

Table 5. Usability and practicality of reviewed passenger-centric metrics in PT systems.

Metric	Implementation	Feasibility	Actionability
PP_2	Detailed AFC data, moderate complexity, simple integration.	Monitoring accuracy, scalable and periodic monitoring.	Simple visualization, easy interpretation.
EW_p	Passenger arrival rates, moderate integration.	Robust monitoring of waiting times.	Actionable insights on waiting times
EWT_s	Simple calculations and easy integration.	Periodic off-peak and real-time monitoring, scalable.	Easy interpretation, Actionable insights on off-peak waiting time.
$ERBT$	Detailed a/typical travel time data, moderate complexity, integrates well (AFC in place).	Recurrent/exceptional delays, periodic monitoring, less scalable(maintenance).	Actionable insight on delay causes, complex visualizations and interpretation.
SD	Detailed travel time data, high complexity, integrates well (if AFC data is used).	Travel time deviations, monitoring details, less scalable (calculations), periodic monitoring.	Complex visualizations and interpretation.
EJT	Detailed AFC journey data, simple calculations and easy integration.	Journey time, scalable and periodic monitoring.	Insights on service quality from both demand and supply perspectives.
$JTBI$	Moderate complexity and easy integration.	Scalable, periodic and real-time monitoring.	Combined and actionable insights on waiting/travel times during off/peak hours.

In terms of ease of implementation, PP_2 is an improved metric but requires detailed AFC data and involves moderately complex calculations to infer the route of the passenger trip based on AFC information. Thus, the metric integrates well with AFC systems already in place. Focusing on passengers' waiting time, EW_p requires accurate data on passengers' arrival rates which is not always possible or easy to collect, e.g., bus passengers, adding moderate integration needs. EWT_s is another improved metric for monitoring off-peak waiting times with very low complexity and easy integration with the existing system for monitoring actual departure times at stops/stations. Similar to PP_2, $ERBT$ also requires detailed passenger travel time data during typical/exceptional traffic conditions, e.g., based on detailed AFC data, which involves more complex calculations to infer the travel times of the different passenger trips data. The metric can integrate well if AFC data is used. SD is a different (ratio) and more complex metric to calculate but has similar data requirements as $ERBT$. Defined as the passenger excess journey time compared to the scheduled one, EJT can be simply calculated for an OD trip given detailed AFC data.

Although $JTBI$ requires some calculations, it involves simple formulas and can easily integrate with existing AVL systems.

As for the practical and operational feasibility, PP_2 provides more accurate monitoring compared to PP_1 and allows for periodic and scalable monitoring with existing AVL and AFC systems. With its higher data requirements (i.e., passenger arrival rates), EW_p can provide more robust monitoring of passenger waiting time over longer periods with variable passenger demands. Focusing on off-peak periods, EWT_s is a scalable metric for both period and real-time monitoring. Practical to monitor different types of delays (recurrent/exceptional), $ERBT$ allows for periodic monitoring but is less scalable as it requires some manual control and maintenance. Similarly, SD also requires more calculations but offers more detailed monitoring of passengers' travel times. A more scalable variant is EJT as it allows for simple monitoring of passenger excess journey times. facilitates directly actionable insights with clear visualizations, supporting periodic monitoring efficiently. $JTBI$ only requires basic AVL data and can therefore allow for scalable and real-time monitoring of both passenger waiting and travel times.

The actionability and interpretability of the reviewed metrics ensure that the insights they provide can be effectively used by PT stakeholders. For instance, PP_2 offers simple visualizations and is easy to interpret, making it straightforward for both passengers and PT operators/agencies to use, e.g., for the latter to improve their service quality. EW_p provides actionable insights specifically focused on passenger waiting times, which can guide targeted interventions to reduce waiting times at important stops/stations. EWT_s is particularly useful for off-peak periods, offering more insights on how to improve service quality during these periods. $ERBT$, while more complex, provides valuable insights into the causes of delays, distinguishing between recurrent and exceptional delays, though its visualizations and interpretations may be more challenging for passengers to understand. SD delivers detailed insights into travel time deviations, but its complexity in both visualization and interpretation requires more analysis. EJT offers comprehensive insights into service quality from both demand and supply perspectives, providing ideas that can inform strategic decisions. Finally, $JTBI$ is useful to gain actionable insights on both waiting and travel times of PT passengers, either during peak or off-peak hours.

3.4. Summary and selected measures

Based on the qualitative assessment of various passenger-centric metrics for use in PT systems, we summarize the analysis by grouping the metrics into different generations based on their data requirements, calculation complexity, and the type of insights they provide. The objective is to identify possible metrics for further quantitative analysis in the next section based on a real-world case study.

Given limited data availability and the project time constraints, some of the reviewed metrics are selected for the quantitative assessment based on criteria or characteristics of the assessed metrics, e.g., data requirements and PT services, see the summary in **Table 6**. Additionally, passenger journey times and types of calculation are also included in the summary table. As indicated in the table, metrics that are between parenthesis are short-

listed for further quantitative analyses in the case study in section 4. Different metrics have been selected based on one or a combination of characteristics that are specific to the metric and the case study.

Besides OTP, Table 6 includes only improved versions of the reviewed metrics, i.e., metrics that are the improvement of other reviewed metrics. OTP is included as it is a traditional metric which is commonly used in practice. For this, OTP is also chosen as a reference metric in the case study.

In addition to OTP, the case study includes the PP_2 metric since it is a passenger-weighted OTP variant, i.e., a passenger punctuality metric, that captures both passengers waiting and travel times. Another reason for selecting this metric is that it can be used for PT services during both peak and off-peak periods. For this same reason and simpler calculations, EJT is also selected. However, it focuses more on passenger travel time, which is important during off-peak hours, and brings a different perspective from OTP and PP_2.

Since the case study includes separate analyses for peak and off-peak hours, EW_p is selected and tested during peak hours as it is more suitable during this period. It also requires simpler calculations and brings a complementary perspective to the other selected metrics as it focuses on passenger waiting times.

Table 6. Summary of some assessed metrics including characteristics that are used as selection criteria for the case study.

Metric (selected)	Data requirement		Passenger journey		PT service			Calculations		
	AVL	AFC (& APC)	Waiting time	Travel time	Long headways	Short headways	Bus	Average	Distributions	Ratio or index
(OTP)	X				X	X				X
(PP_2)	X	X	X	X	X	X				X
(EW_p)	X	X	X			X				X
EWT_s	X		X		X			X	X	
$ERBT$	X	X		X	X	X	X		X	
SD	X	X		X	X	X	X	X	X	X
(EJT)	X	X		X	X	X	X		X	
$JTBI$	X		X	X	X	X	X	X	X	X

X if the metric has the characteristic, (.) if a metric is selected for the quantitative assessment.

4. Case study: Stockholm commuter rail

To quantitatively assess and compare the selected metrics, Stockholm commuter rail during (winter) 2015 is chosen as a case study. For the sake of simplicity, the assessment is performed on a specific line of the network, namely Bål-Nyh, see Figure 10.

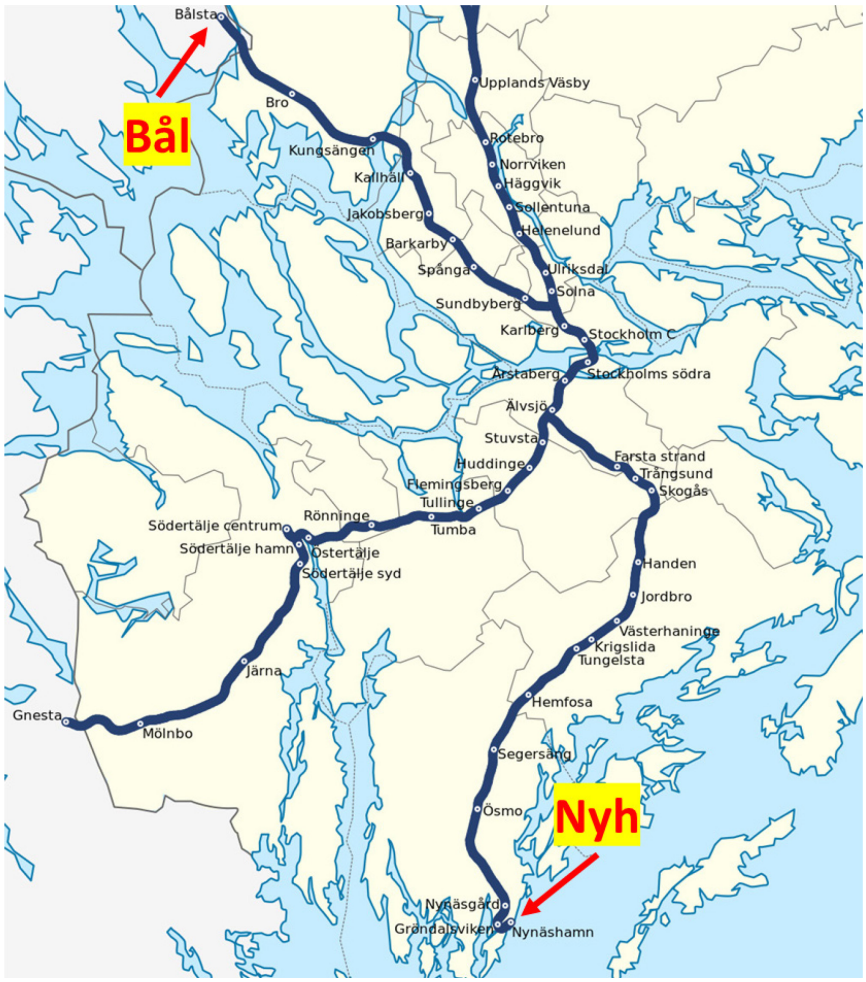


Figure 10. Map of Stockholm commuter rail including the line between Bål and Nyh which is the focus of the case study (Frohne et al., 2014).

First, the input data is described including different preprocessings and assumptions. Quantitative analyses and comparisons are performed and presented at line and station levels for different periods of the day, i.e., peak and off-peak hours. Finally, the assessment results and insights are discussed.

4.1. Input data and preprocessing

This subsection details the data sets used in the case study, focusing on two primary sources, namely AFC-based passenger demand estimates and AVL-based traffic data. The preprocessing steps and key assumptions applied to these data sets are also discussed.

4.1.1. Passenger demand estimates

The passenger demand estimates that are used in this case study are derived from a sample of AFC or smart card data (between week 38 and week 42 in 2015) about passenger boardings at different stations of the studied line and periods. Based on such data, the passenger demand, i.e., OD matrices, can be estimated using different methods to infer the most likely paths taken by passengers, e.g., alighting stations. The passenger demand estimates in this case study are based on the findings of Ait-Ali and Eliasson (2019), employing an entropy-maximization approach to infer alighting stations (Ait-Ali and Eliasson, 2021). The demand estimates are presented in 15-minute intervals and are restricted to a typical workday during the winter of 2015 to capture standard commuting patterns.

The spatial and temporal variations in passenger demand are visualized in **Figure 11** and **Figure 12**, respectively. **Figure 11** illustrates the spatial variation of passenger boardings at different stations on the studied line between *Bål* (in the north) and *Nyh* (in the south). The data clearly shows a concentration of boardings at central stations along the line, indicating e.g., transfers from/to other lines. Additionally, higher boarding numbers are observed at the first stations compared to the last stations in the direction of travel on the studied line.

4.2. Quantitative comparisons

In this subsection, we analyse service punctuality and compare different quantitative measures of passenger punctuality by integrating the AVL and AFC-based data previously presented. The analysis is divided into two main parts: first, an examination of overall punctuality on the studied line; and second, a focus on punctuality during peak and then off-peak periods using selected metrics.

4.2.1. Overall passenger punctuality

Analysing the overall punctuality is an important initial step for evaluating the reliability of the commuter rail service. In this case study, we focus on a specific line, analysing punctuality across different periods including both peak and off-peak hours. The key metrics used in this analysis are OTP and PP2, which are calculated and compared using various delay thresholds.

OTP and delay thresholds

As stated earlier, OTP is typically calculated with a 3-minute threshold for commuter services as in our case study. In this analysis, we examine how the OTP score, measured at different time periods of the day, varies across different delay thresholds. The results are presented in Figure 14, which shows the OTP scores for various thresholds and time periods. The stepwise shape of the curves in the figure is due to the limited time precision of traffic data, i.e., up to 1 minute. All earlier arrivals are considered on time, i.e., within the delay threshold.

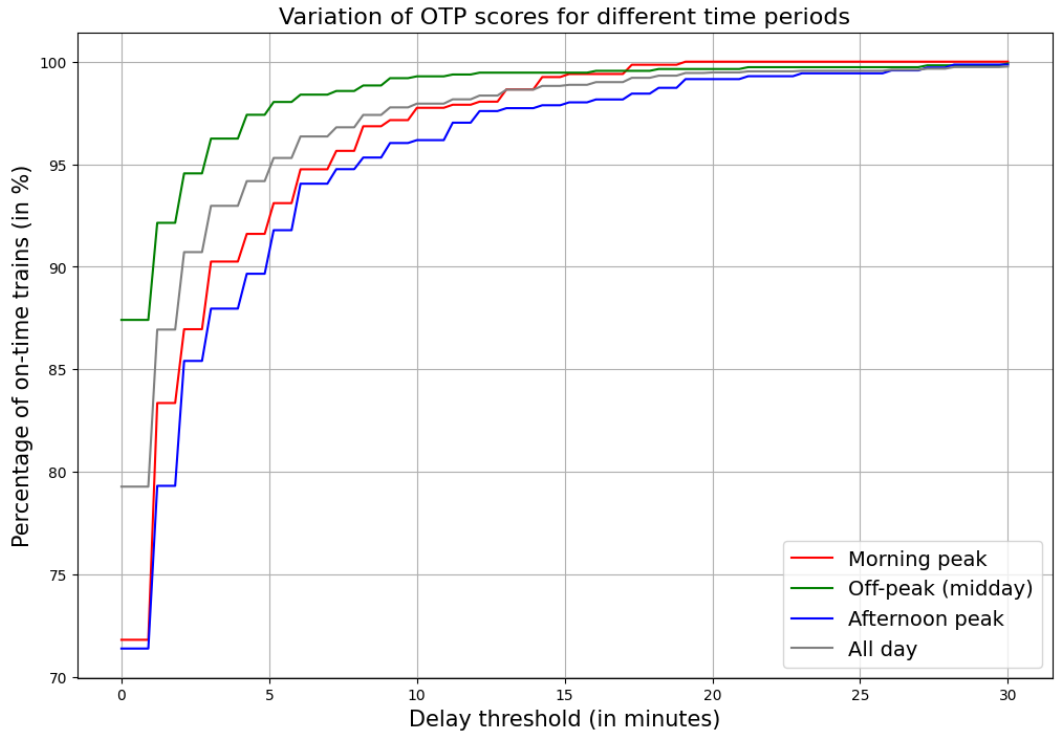
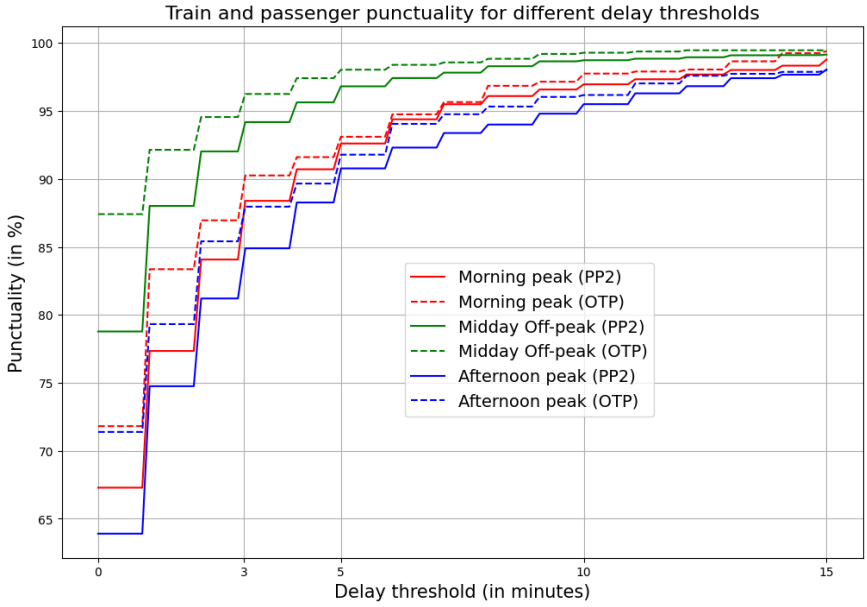
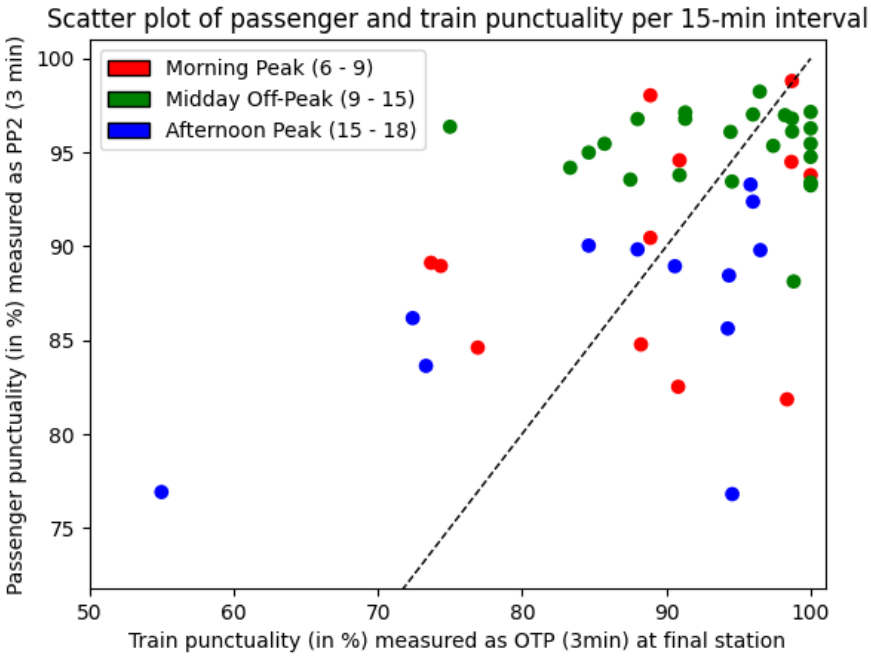


Figure 14. Overall punctuality measured as OTP for different delay thresholds and time periods of the day.



(a)



(b)

Figure 16. Comparison of train and passenger punctuality scores **(a)** for different delay thresholds, and **(b)** for a 3-min threshold in 15-min time intervals.

This initial comparative analysis highlights the importance of considering passenger-centric metrics, such as PP2, alongside train-centric metrics when evaluating the overall punctuality of commuter rail services. Passenger metrics such as PP2 can better capture passenger experience, particularly during the different travel periods. Since disaggregate

analyses show more nuanced results, the following section further analyse passenger punctuality by using other selected metrics and focusing on two different periods of the day, namely peak and off-peak hours.

4.2.2. Passenger punctuality during off-peak hours

We first focus on analysing passenger punctuality during off-peak hours using the selected EJT metric, which is suited for long-headway services typical of off-peak hours. The EJT metric is compared to the passenger punctuality PP2. As discussed in the literature, the calculation of EJT is based on the distribution of actual travel times relative to the average scheduled times for each origin-destination (OD) pair.

Figure 17 illustrates the distribution of actual travel time percentiles compared to the average scheduled time for all trips on the studied line during off-peak hours. The data reveals that the average scheduled travel time across the line is approximately 34 minutes. However, half of the trips (50^{th} percentile) have an actual travel time of around 28 minutes, while the 90^{th} percentile of trips has up to an hour and 10 minutes in actual travel time.

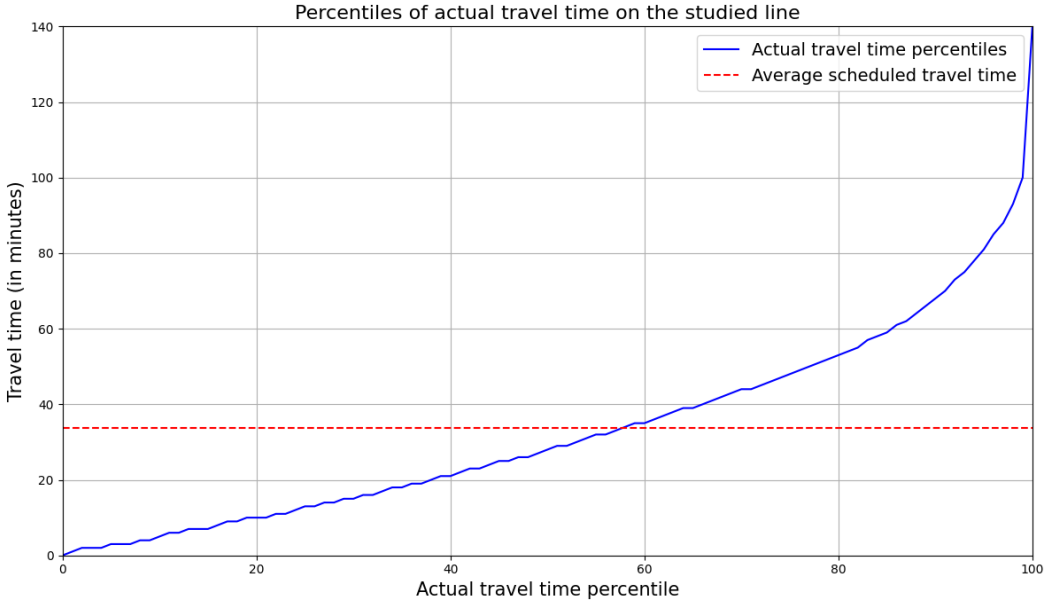
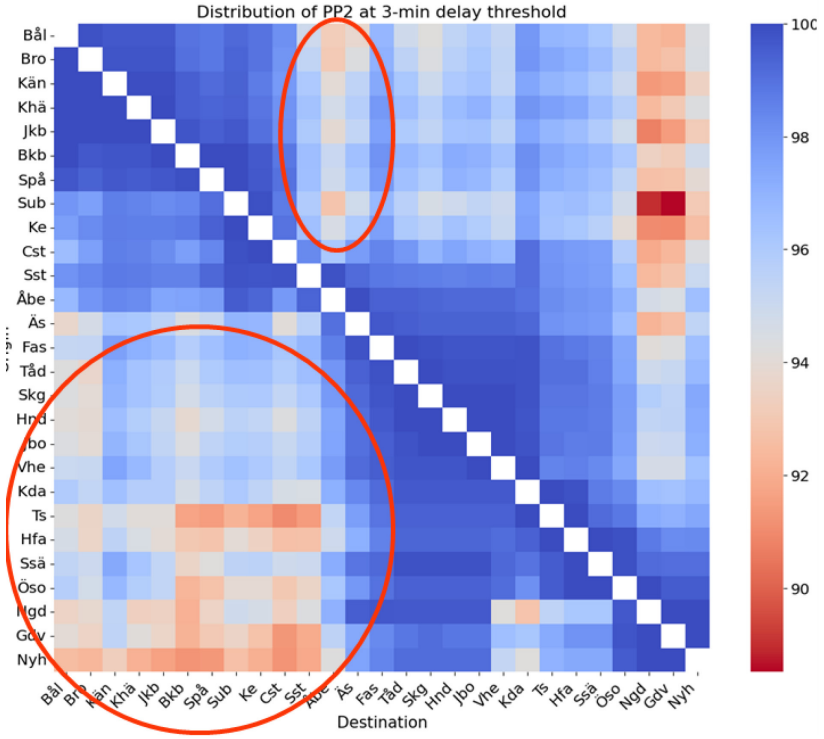
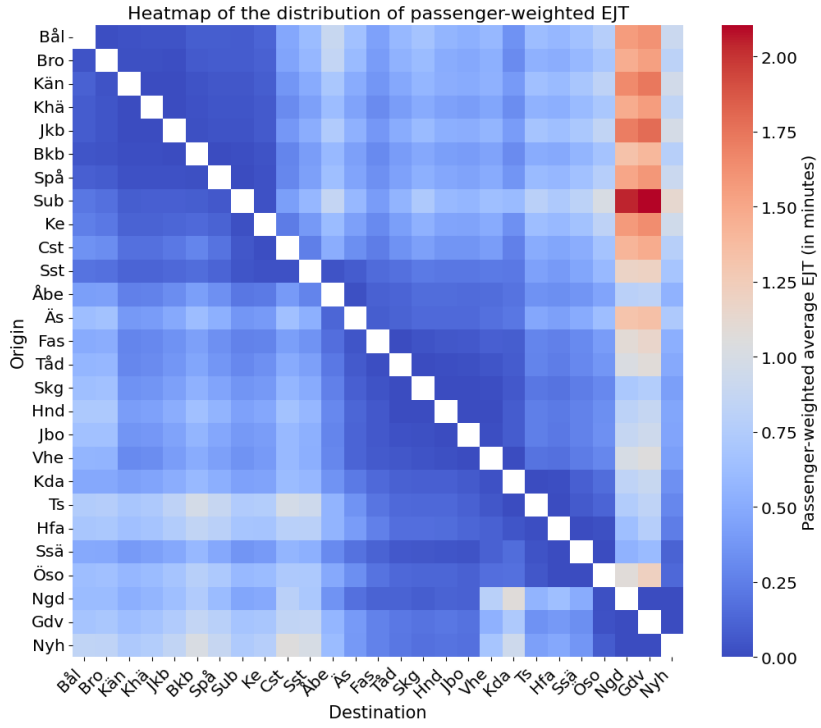


Figure 17. Distribution of actual travel time compared to the average scheduled time on the studied line during off-peak hours.

Using the distributions for each OD pair, combined with passenger ridership data, the passenger-weighted EJT is calculated. **Figure 18** provides a visual comparison of punctuality scores across different OD pairs on the studied line. **Figure 18**(a) shows the PP2 scores calculated with a 3-minute delay threshold, while **Figure 18**(b) presents the passenger-weighted EJT scores. In both figures, lower punctuality scores are in red, indicating areas with more significant punctuality issues, whereas higher scores are shown in blue.



(a)



(b)

Figure 18. Spatial distribution of **(a)** PP2 scores at 3-min threshold and **(b)** passenger-weighted EJT scores during off-peak hours.

The comparison between *EJT* and PP2 reveals that both metrics generally identify similar OD pairs with punctuality issues. However, there are differences in the specific pairs highlighted as least punctual, see the red-encircled regions in **Figure 18**(a). PP2 highlights many northbound pairs where a significant number of passengers arrive more than 3 minutes late. On the other hand, *EJT* reflects a different aspect of the passenger experience by highlighting the (southbound) pairs (in dark red in **Figure 18** (a) and (b)) where the total travel time deviates more significantly from the scheduled time.

To summarize the results, Table 7 presents the punctuality scores for the entire line during off-peak hours. The OTP score, measured with a 3-minute delay threshold, is 96.3%, which is slightly higher than the PP2 score at 94.2%. The overall *EJT* score shows an average excess travel time of around 19 seconds per passenger.

Table 7. Summary of the scores for tested punctuality metrics during off-peak hours.

Punctuality metric	Score (unit)
OTP at 3-min delay threshold	96.3%
PP2 at 3-min delay threshold	94.2%
Passenger-weighted average EJT	19 seconds per passenger

Since *EJT* is an absolute metric expressed in minutes per passenger, it cannot be directly compared to the binary/percentage-based scores like OTP and PP2. However, the *EJT* score is more flexible as it has the advantage of not requiring a predefined delay threshold for its calculations. Besides, it is particularly suitable for services with long headways during off-peak hours. Although *EJT* cannot be compared to other binary metrics like PP2, it is possible to explore their correlation. **Figure 19** presents a scatter plot showing the relationship between PP2 and *EJT* scores across different OD pairs during off-peak hours.

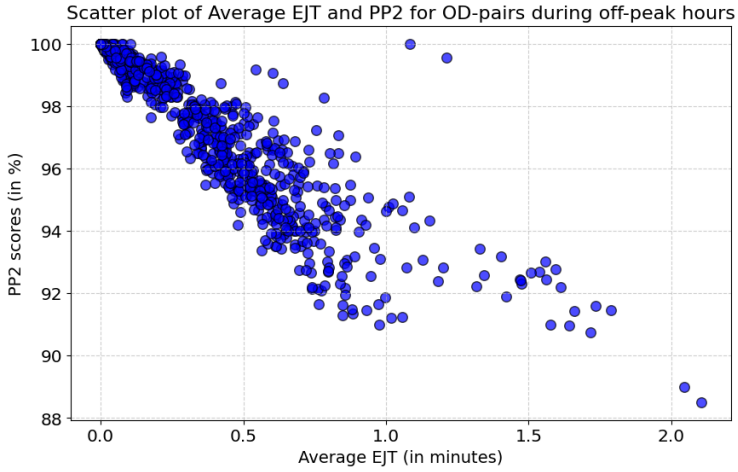


Figure 19. Scatter plot showing the correlation between EJT and PP2 scores during off-peak hours.

The scatter plot in **Figure 19** indicates a correlation between these two metrics: higher *EJT* scores (longer excess journey times for passengers) are associated with lower PP2 scores (fewer passengers arrived within the promised arrival time), and vice versa. The

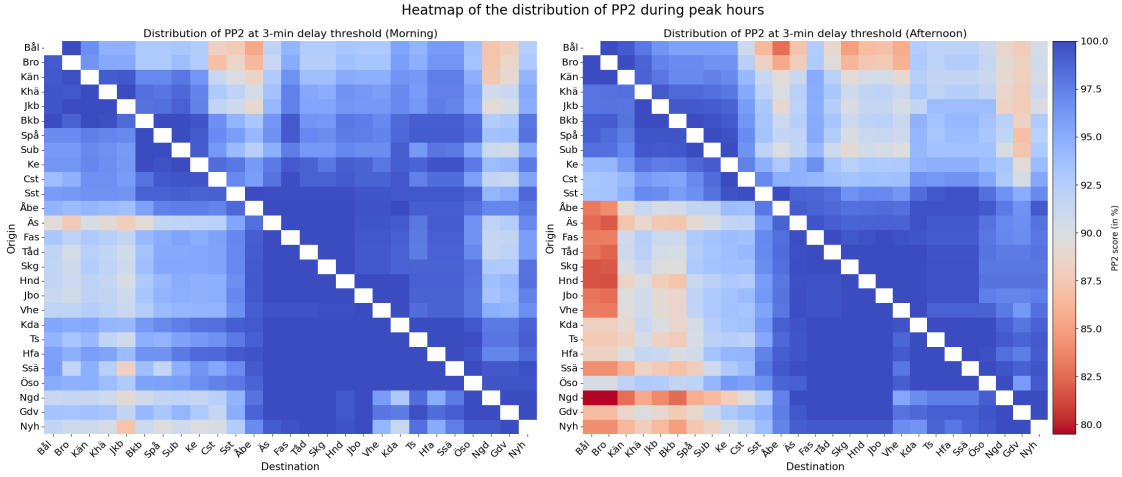
correlation is particularly strong for lower EJT values and higher PP2 scores. Higher values of EJT, which are associated with lower PP2 scores, show more variance.

These findings show that both passenger metrics capture aspects of passenger punctuality but from different perspectives. EJT is useful in identifying areas where total travel time is significantly impacted, while PP2 highlights delays that affect many passengers, even if those delays do not consistently prolong the total travel time. In the following section, we will extend this analysis to cover peak-hour periods, examining how passenger punctuality varies with an additional metric focusing on passengers' waiting time, i.e., expected (passenger-weighted) waiting time (EW_p).

4.2.3. Passenger punctuality during peak hours

The focus here is on analysing passenger punctuality during peak hours, specifically during the morning and afternoon periods, using previously tested metrics (i.e., OTP, PP2, and EJT) and EW_p focusing on expected waiting times.

We first compare PP2 (3-min threshold) results between the morning and afternoon peak hours. **Figure 20** illustrates the spatial distribution of PP2 scores across all OD pairs for both time periods, with **Figure 20 (a)** and **(b)** showing the morning and afternoon peak hours, respectively. A common colour coding is used across both figures for direct comparison.



(a) (b)

Figure 20. Spatial distribution of PP2 scores during **(a)** morning and **(b)** afternoon peak hours.

The results indicate that PP2 scores are generally lower during the afternoon peak hours, as indicated by more dark red areas in **Figure 20(b)**. Regions with punctuality issues (lower PP2 scores, marked in red) that are highlighted in the morning period are also present in the afternoon, but the extent and severity of these issues are more pronounced in the afternoon, with larger and darker areas indicating a greater number passenger and more OD pairs experiencing more significant delays. Notably, afternoon peak hours show more punctuality issues in northbound OD pairs, as seen in the lower half of **Figure 20(b)**.

As for off-peak hours, we also calculate the EJT metric, for both morning and afternoon peak hours, from the distributions of actual travel time percentiles compared to the scheduled travel time for each OD pair. **Figure 21(a)** and **(b)** presents the spatial distribution of EJT scores during morning and afternoon peak hours, respectively. Again, a common colour bar is used for direct comparison between the two periods.

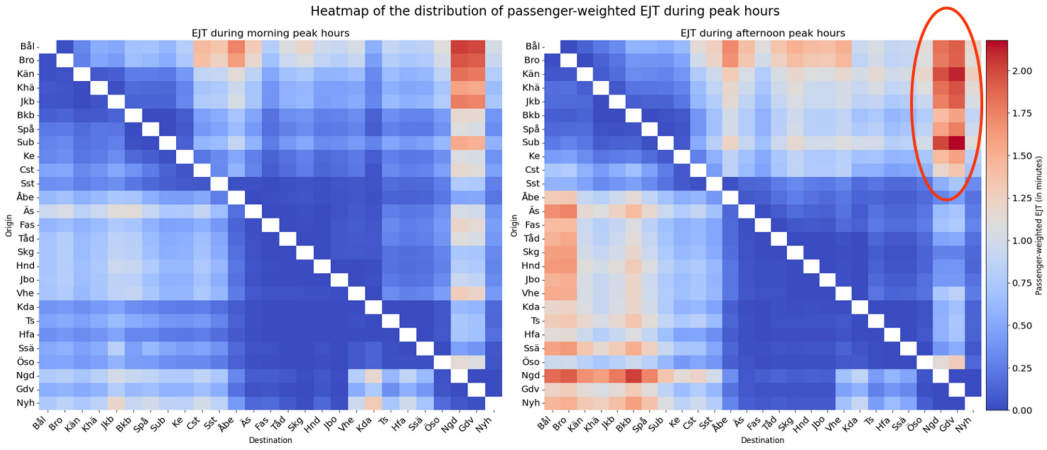
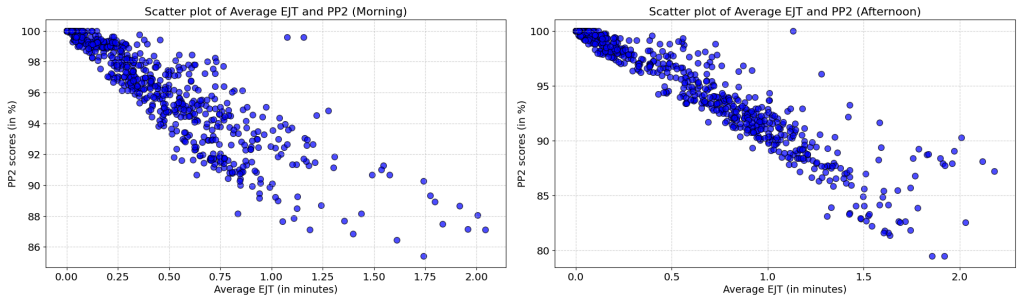


Figure 21. Spatial the distribution of passenger-weighted EJT during **(a)** morning and **(b)** afternoon peak hours.

Higher EJT scores, in **Figure 21**, highlight areas with punctuality issues, i.e., red areas which indicate significant deviations from scheduled travel times. However, while there are similarities in the regions pinpointed by both metrics, there are some differences, , see the red-encircled regions in **Figure 21(b)**. For instance, EJT highlights certain southbound OD pairs with extreme punctuality issues during the afternoon peak hours, which were less prominent in the PP2 analysis in **Figure 20**. These differences are due to the way EJT captures the overall travel time deviation without relying on a fixed delay threshold, unlike PP2 which uses a 3-minute threshold that may exclude some (consistent) travel time deviations.

To explore the relationship between PP2 and EJT, a scatter plot analysis is performed for both morning and afternoon peak hours, as shown in **Figure 22 (a)** and **(b)**. As for off-peak hours, scatter plots indicate a negative correlation between the scores in both periods. The correlation is stronger and less scattered for the afternoon period and for OD pairs with better punctuality.



(a) (b)

Figure 22. Scatter plotting showing the correlation between EJT and PP2 for **(a)** morning and **(b)** afternoon peak hours.

In addition to PP2 and EJT, we also analyse the EW or rather the expected (passenger-weighted) waiting time (noted EW_p earlier). It is particularly useful during peak periods when high passenger volumes and short-headway services make waiting time a more critical component of the overall journey experience. Lower EW scores indicate better passenger punctuality and more regular services, making it a valuable metric for understanding the impact of service irregularities on passengers.

Figure 23 shows the resulting EW scores for different departure stations along the studied line during morning and afternoon peak hours. The EW scores range between 8 to 18 minutes for both periods, with stations close to terminal stations having generally higher EW due to lower service frequency compared to more central stations. Some central stations, despite having more frequent services, also show slightly higher EW scores due to higher passenger ridership, reflecting the impact of both service supply and passenger demand on waiting times. Thus, the EW metric can highlight the interplay between passenger ridership, service frequency, and regularity during peak hours, which can complement the other metrics.

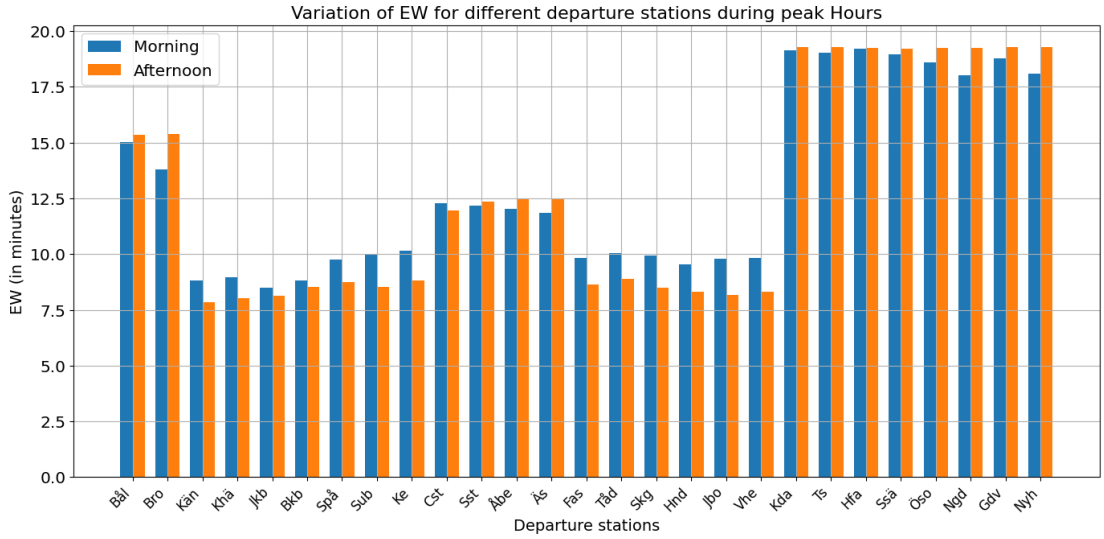


Figure 23. EW scores for different departure stations in the studied line during peak hours.

Table 8 summarizes the aggregated score of the studied punctuality metrics for the morning and afternoon peak hours. The aggregated results show that punctuality scores are generally better during the morning peak compared to the afternoon, except for EW, indicating a slightly higher passenger waiting time in the afternoon peak hours. Note that EW is theoretically a more suitable metric to use during peak hours.

Table 8. Summary of the calculated punctuality metrics during peak hours.

Metric (unit)	Morning	Afternoon
OTP at 3-min delay threshold (in percent)	90.3%	88.0%
PP2 at 3-min delay threshold (in percent)	88.4%	84.9%
EJT (in seconds per passenger)	27	31
EW (in minutes)	13	12.7

The comparison of these metrics highlights the importance of using a various measure to gain a more comprehensive understanding of punctuality during peak hours. While metrics, e.g., OTP and PP2, provide a general overview of punctuality based on delay thresholds, EJT can offer more insights of travel time deviations, and EW can complement it with passenger waiting times analysis due to service irregularities. The differences observed between these metrics underscore the need for a nuanced approach when assessing punctuality, as different metrics capture different aspects of the passenger experience, often leading to varying conclusions. It is therefore essential to continue exploring and comparing various existing metrics in practice across different scenarios to better understand how they can be combined for a comprehensive assessment of passenger punctuality in PT systems.

5. Concluding remarks

The study explores different passenger-centric punctuality metrics to address the limitations of traditional vehicle-centric metrics that are commonly used in PT systems. Both quantitative and qualitative insights are discussed with representatives from different PT actors, namely *Skånetrafiken*, the regional public transport authority (PTA) in the Scania region, and *Transportstyrelsen*, the national transport regulator in Sweden.

In this chapter, the importance of the reviewed metrics is highlighted along with practical recommendations for improving PT service monitoring and management. It also summarises some of the challenges and potentials of these metrics as well as directions for future works.

5.1. Discussions and insights

The choice of delay thresholds plays a crucial role when using OTP and its variants, including passenger-centric ones, e.g., passenger-weighted OTP (PP2). PT stakeholders, including operators and PTAs, need to be aware of how sensitive the punctuality scores are to this parameter. A sensitivity analysis, as shown in the quantitative assessment (see 4.2.1), is one possible method to understand the influence of different thresholds on the punctuality results.

As expected, the quantitative analysis reveals lower punctuality during morning and afternoon peak hours, both from a vehicle and passenger perspective. PTAs and other PT stakeholders should complement overall punctuality assessments with period-specific analyses, e.g., peak versus off-peak. The qualitative review also suggests that numerous metrics are more suitable for these specific periods. Additionally, the case study illustrates a correlation between passenger ridership and vehicle delays, highlighting the need for punctuality analyses from a passenger perspective, especially when doing cross-comparisons between periods and locations with large differences in ridership.

Passenger-centric metrics, as demonstrated in the case study, provide a more in-depth understanding of punctuality compared to vehicle-centric metrics. However, these metrics require good-quality demand data. Additionally, there are some challenges in collecting such (good quality) data, such as the PT vehicle fleet not being fully equipped with APC, inconsistent ticket validation by passengers, missing alighting data, and varying data quality across transport modes (e.g., buses typically have better demand data than trains). Based on the available data, some metrics require only aggregate demand estimates (e.g., passenger-weighted OTP), while others need more detailed data.

The case study further illustrates that traditional vehicle-centric metrics tend to overestimate passenger punctuality, especially during peak hours. As more automatic data collection systems are implemented and demand data quality improves, PTAs and

PT operators could consider updating their overarching punctuality goals to reflect passenger-centric metrics, as these are more closely linked to passenger satisfaction. This shift could also inform new requirements in tendering contracts with PT operators and be useful in monitoring existing contracts, which are currently mainly evaluated using vehicle-centric metrics. With improved data collection systems, there is also potential for real-time monitoring and operational management using passenger-centric metrics to improve the resilience of PT systems and their service reliability during disruptions.

5.2. Summary of challenges and potentials

While the reviewed metrics offer valuable insights into passenger punctuality, several challenges have been identified. One major challenge is data availability and quality, particularly for metrics which depend on more detailed AFC data. Thus, ensuring high-quality and consistent data collection across PT modes and systems is important for effective implementation. Integration with existing automatic collection systems can also pose difficulties as well as the deployment and maintenance of such systems.

Another related challenge is the complexity of certain metrics, which may require more data processing, e.g., estimating alightings, and can therefore be resource-intensive when applied to large networks or used for real-time monitoring. Furthermore, certain metrics can be difficult to interpret and communicate effectively to various PT stakeholders.

Despite these challenges, there are several potentials that can be highlighted:

- With enhanced data collection systems, more accurate insights can be gained into where and when passenger delays occur. This allows for targeted improvements, such as adjusting schedules or optimizing routes during peak demand periods.
- When combined with real-time data, passenger-centric metrics can improve operational management based on real-time accurate monitoring of passenger punctuality.
- Passenger-centric metrics allow PT operators to better align their service performance targets with passenger needs. This is also useful for PTAs when revising tendering contracts with operators, ensuring that passenger punctuality and satisfaction become a more central focus of performance evaluations.
- Passenger-centric metrics can potentially help identify underserved areas and periods where passengers face disproportionate delays. This can lead to more equitable service improvements.
- As the availability of high-quality data improves, the same metrics can be applied across different PT modes, allowing for a scalable and more integrated evaluation of service reliability across multiple modes.

5.3. Directions for future works

This study focused primarily on punctuality (and regularity to a lesser extent) as key aspects of PT reliability. Future research can build upon this by exploring broader

dimensions of service reliability and by applying these passenger-centric metrics across different PT modes. What follows is a number of directions for ture work:

- Given that the case study was focused on commuter rail, future research could study bus services, which have their own unique dynamics and challenges. Buses often have better data quality, making them an ideal subject for further quantitative experimentation by investigating bus-specific metrics. Furthermore, such studies could extend passenger-centric metrics to capture multimodal journeys, integrating data across multiple lines and modes. By including the effects of transfers and overall journey times, this would provide a better understanding of the passenger experience, compared to focusing on single modes or routes.
- An important direction for future research is to investigate the updating of overarching punctuality goals and performance regimes in procurement contracts. As the understanding of passenger-centric punctuality improves, there is an opportunity to revise how these performance measures are defined by the PTAs and enforced in tendering contracts with PT operators. Ensuring that punctuality metrics align with passenger needs, rather than vehicle-centric goals, can incentivize operators to focus on delivering a higher quality of service and increase passenger satisfaction.
- Future work could also explore broader resilience aspects of PT systems, particularly how they respond to disruptions and delays. Sensitivity analyses can help explore how demand variations and different types of delays (e.g., minor versus severe) impact passenger punctuality.
- Additional research could focus on refining delay classifications (type, length and frequency) and examining their non-linear impacts on passenger satisfaction. Understanding which types of delays (e.g., short vs. long, expected vs. unexpected) affect passengers the most can inform more targeted service interventions.
- Traffic information systems and how they influence passenger punctuality should also be studied further. Investigating the role of real-time information in enhancing passenger punctuality and satisfaction can offer insights into how such systems can be better integrated into PT operations.

References

Ait-Ali, A. & Eliasson, J. 2019. Dynamic Origin-Destination Estimation Using Smart Card Data: An Entropy Maximisation Approach. *Preprint arXiv:1909.02826*.

Ait-Ali, A. & Eliasson, J. 2021. The value of additional data for public transport origin–destination matrix estimation. *Public Transport*.

Ait Ali, A., Eliasson, J. & Warg, J. 2022. Are commuter train timetables consistent with passengers' valuations of waiting times and in-vehicle crowding? *Transport Policy,* 116**,** 188-198.

Bagherian, M., Cats, O., van Oort, N. & Hickman, M. Measuring passenger travel time reliability using smart card data. TRISTAN IX: Triennial Symposium on Transportation Analysis, Oranjestad, Aruba, 2016.

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A. & Zou, B. 2010. Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States.

Barabino, B., Di Francesco, M. & Mozzoni, S. 2015. Rethinking bus punctuality by integrating Automatic Vehicle Location data and passenger patterns. *Transportation Research Part A: Policy and Practice,* 75**,** 84-95.

Beirão, G. & Sarsfield Cabral, J. A. 2007. Understanding attitudes towards public transport and private car: A qualitative study. *Transport Policy,* 14**,** 478-489.

Berggren, U., D'Agostino, C., Svensson, H. & Brundell-Freij, K. 2022. Intrapersonal variability in public transport path choice due to changes in service reliability. *Transportation,* 49**,** 1517-1547.

Blayac, T. & Stéphan, M. 2021. Are retrospective rail punctuality indicators useful? Evidence from users perceptions. *Transportation Research Part A: Policy and Practice,* 146**,** 193-213.

Börjesson, M. & Eliasson, J. 2011. On the use of "average delay" as a measure of train reliability. *Transportation Research Part A: Policy and Practice,* 45**,** 171-184.

CEN 2002. Transportation - Logistics and Service - Public Passenger Transport - Service Quality Definition, Targeting and Measurement. EU.

Chan, J. 2007. *Rail transit OD matrix estimation and journey time reliability metrics using automated fare data.* Massachusetts Institute of Technology.

Cramer, A., Cucarese, J., Tran, M., Lu, A. & Reddy, A. 2009. Performance Measurements on Mass Transit:Case Study of New York City Transit Authority. *Transportation Research Record,* 2111**,** 125-138.

Danaher, A., Wensley, J., Dunham, A., Orosz, T., Avery, R., Cobb, K., Watkins, K., Queen, C., Berrebi, S. & Connor, M. 2020. Minutes Matter: A Bus Transit Service Reliability Guidebook.

dell'Olio, L., Ibeas, A. & Cecin, P. 2011. The quality of service desired by public transport users. *Transport Policy,* 18**,** 217-227.

den Heijer, A. 2018. *Passenger punctuality: Assessing the impact of disruptions.* master thesis, Delft University of Technology.

Denti, E. & Burroni, L. 2023. Delay Indices for Train Punctuality. *Information,* 14**,** 269.

Diab, E. I., Badami, M. G. & El-Geneidy, A. M. 2015. Bus Transit Service Reliability and Improvement Strategies: Integrating the Perspectives of Passengers and Transit Agencies in North America. *Transport Reviews,* 35**,** 292-328.

Ferreira, L. & Higgins, A. 1996. Modeling reliability of train arrival times. *Journal of transportation engineering,* 122**,** 414-420.

Friman, M. 2004. Implementing quality improvements in public transport. *Journal of Public transportation,* 7**,** 49-65.

Frohne, E., Kildor, Tomiwoj & Schröter, U. 2014. Linjekarta för Stockholms pendeltåg. Wikimedia Commons.

Furth, P. G. & Muller, T. H. J. 2007. Service Reliability and Optimal Running Time Schedules. *Transportation Research Record,* 2034**,** 55-61.

Ghofrani, F., He, Q., Goverde, R. M. P. & Liu, X. 2018. Recent applications of big data analytics in railway transportation systems: A survey. *Transportation Research Part C: Emerging Technologies,* 90**,** 226-246.

Gittens, A. & Shalaby, A. 2015. Evaluation of Bus Reliability Measures and Development of a New Composite Indicator. *Transportation Research Record,* 2533**,** 91-99.

Hendren, P., Antos, J., Carney, Y. & Harcum, R. 2015. Transit Travel Time Reliability: Shifting the Focus from Vehicles to Customers. *Transportation Research Record,* 2535**,** 35-44.

Joborn, M. & Ranjbar, Z. 2022. Understanding causes of unpunctual trains: Delay contribution and critical disturbances. *Journal of Rail Transport Planning & Management,* 23**,** 100339.

Kristoffersson, I. & Pyddoke, R. A traveller perspective on railway punctuality: Passenger loads and punctuality for regional trains in Sweden. RailNorrköping 2019. 8th International Conference on Railway Operations Modelling and Analysis (ICROMA), Norrköping, Sweden, June 17th–20th, 2019, 2019. Linköping University Electronic Press, 565-578.

Lee, A., van Oort, N. & van Nes, R. 2014. Service Reliability in a Network Context:Impacts of Synchronizing Schedules in Long Headway Services. *Transportation Research Record,* 2417**,** 18-26.

Mishalani, R. G., McCord, M. M. & Wirtz, J. 2006. Passenger Wait Time Perceptions at Bus Stops: Empirical Results and Impact on Evaluating Real-Time Bus Arrival Information. *Journal of Public Transportation,* 9**,** 89-106.

NASEM 2006. Using Archived AVL-APC Data to Improve Transit Performance and Management. *National Academies of Sciences, Engineering, Medicine*.

Nelldal, B.-L., Andersson, J. & Fröidh, O. 2019. *Utveckling av utbud och priser på järnvägslinjer i Sverige 1990-2019: Avreglering och konkurrens mellan tåg, flyg och buss samt jämförelse mellan tåg-och resenärspunktlighet,* Rapport, Kungliga Tekniska högskolan (KTH).

NetworkRail 2017. Definitions of Railway Performance Metrics.

Nielsen, O. A. 2000. A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B: Methodological,* 34**,** 377-402.

Nielsen, O. A., Landex, O. & Frederiksen, R. D. 2008. Passenger delay models for rail networks. *Schedule-Based Modeling of Transportation Networks: Theory and applications.* Springer.

Parbo, J., Nielsen, O. A. & Prato, C. G. 2016. Passenger Perspectives in Railway Timetabling: A Literature Review. *Transport Reviews,* 36**,** 500-526.

Pelletier, M.-P., Trépanier, M. & Morency, C. 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies,* 19**,** 557-568.

Rudnicki, A. 1997. Measures of Regularity and Punctuality in Public Transport Operation. *IFAC Proceedings Volumes,* 30**,** 661-666.

TCQSM 2013. Transit Capacity and Quality at Service Manual. Third Edition ed. Transportation Research Board (TRB): National Academies of Sciences, Engineering, Medicine.

Tirachini, A., Godachevich, J., Cats, O., Muñoz, J. C. & Soza-Parra, J. 2022. Headway variability in public transport: a review of metrics, determinants, effects for quality of service and control strategies. *Transport Reviews,* 42**,** 337-361.

Trafa. 2023. *Ändrat statistiskt mått för punktlighet på järnväg* [Online]. Available: https://www.trafa.se/sidor/matt-for-punktlighet/ [Accessed].

Trafikverket 2020. En punktligare tågtrafik: sammanställning av Trafikverkets åtgärder 2017–2019. *In:* TRAFIKVERKET (ed.).

Transportstyrelsen 2023. Resenärers syn på järnvägsmarknaden 2023.

Uniman, D. L., Attanucci, J., Mishalani, R. G. & Wilson, N. H. 2010. Service reliability measurement using automated fare card data: application to the London underground. *Transportation research record,* 2143**,** 92-99.

van Loon, R., Rietveld, P. & Brons, M. 2011. Travel-time reliability impacts on railway passenger demand: a revealed preference analysis. *Journal of Transport Geography,* 19**,** 917-925.

van Oort, N. 2016. Incorporating enhanced service reliability of public transport in cost-benefit analyses. *Public Transport,* 8**,** 143-160.

Vanhanen, K. & Kurri, J. 2005. Quality factors in public transport. *Helsinki University of Technology http://transportation. org. il/en/node/3017*.

Wardman, M. 2001. A review of British evidence on time and service quality valuations. *Transportation Research Part E: Logistics and Transportation Review,* 37**,** 107-128.

Wei, Y., Yang, X., Xiao, X., Ma, Z., Zhu, T., Dou, F., Wu, J., Chen, A. & Gao, Z. 2024. Understanding the Resilience of Urban Rail Transit: Concepts, Reviews and Trends. *Engineering*.

Wilson, N. H., Nelson, D., Palmere, A., Grayson, T. H. & Cederquist, C. 1992. Service-quality monitoring for high-frequency transit lines. *Transportation Research Record*.

Wolters, G. 2016. Passenger punctuality: An analysis of the method of calculation and describing models.

Zhao, J., Frumin, M., Wilson, N. & Zhao, Z. 2013. Unified estimator for excess journey time under heterogeneous passenger incidence behavior using smartcard data. *Transportation Research Part C: Emerging Technologies,* 34**,** 70-88.

K2 is Sweden’s national centre for research and education on public transport. This is where academia, the public sector and industry meet to discuss and develop the role of public transport.

We investigate how public transport can contribute to attractive and sustainable metropolitan areas of the future. We educate members of the public transport sector and inform decision-makers to facilitate an educated debate on public transport.

K2 is operated and funded by Lund University, Malmö University and VTI in cooperation with Region Stockholm, Region Västra Götaland and Region Skåne. We receive financial support from Vinnova, Formas and the Swedish Transport Administration.

www.k2centrum.se/en