Aalto University

OPEN ACCESS

This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Roncoli, Claudio; Chandakas, Ektoras; Kaparias, Ioannis

Estimating on-board passenger comfort in public transport vehicles using incomplete automatic passenger counting data

Published in:
Transportation Research Part C: Emerging Technologies

DOI:
10.1016/j.trc.2022.103963

Published: 01/01/2023

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Roncoli, C., Chandakas, E., & Kaparias, I. (2023). Estimating on-board passenger comfort in public transport vehicles using incomplete automatic passenger counting data. *Transportation Research Part C: Emerging Technologies*, *146*, Article 103963. https://doi.org/10.1016/j.trc.2022.103963

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Transportation Research Part C 146 (2023) 103963

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Estimating on-board passenger comfort in public transport vehicles using incomplete automatic passenger counting data

Claudio Roncoli [a], Ektoras Chandakas [b], Ioannis Kaparias [c,*]

[a] Department of Built Environment, Aalto University, Espoo, Finland
[b] LVMT UMR-T 9403, Ecole des Ponts, Université Gustave Eiffel, Champs-sur-Marne, France
[c] Transportation Research Group, University of Southampton, UK

A B S T R A C T

The prevention of crowding inside buses, trams and trains is an important component of on-board passenger comfort and is central to the provision of good public transport services. In light of the COVID-19 pandemic and the associated significant reduction in public transport patronage and, more importantly, in passenger confidence, the avoidance of crowds by passengers and operators alike becomes even more critical. This is where the provision of information on on-board comfort becomes a necessity. The present study, therefore, proposes a new Kalman filter based estimation scheme for on-board comfort levels, employing historical and current (same-day) non-exhaustive Automatic Passenger Counting data, as well as Automatic Vehicle Locating measurements. The accuracy and reliability of the estimation is, then, evaluated through application to the tramway network of the French city of Nantes. The results suggest that the proposed method is able to deliver good estimation accuracy, both in terms of absolute passenger numbers, but also, more crucially, in terms of on-board comfort Levels of Service.

1. Introduction

The number of passengers on board a public transport vehicle is a prominent constituent factor of on-board passenger comfort and is, therefore, critical for both operators and passengers. It plays an important role in implementing control strategies and improving schedule adherence and is also a key determinant of the quality of service. In addition, knowledge about current and anticipated on-board volumes is key to preventing crowding and ensuring effective observation of social distancing in the post-COVID-19 reality. Indeed, keeping a social distance of 1 or 2 m at stops and stations, and, more crucially, on-board public transport vehicles, is likely to remain desirable by passengers, regardless of whether it is a legal or advisory requirement. Early evidence on public transport occupancy levels from cities around the world from the onset of the COVID-19 pandemic, unfortunately, shows that travellers' confidence towards public transport may have been significantly dented (Transport Focus, 2021). And while it can be expected that mass vaccination and the decreasing virulence of the disease over time will restore some of this confidence in the long-run, it looks likely that much of the damage may be irreparable in the short- to medium-term, and that it may be a long time until passengers feel again comfortable travelling on public transport systems operating at or near capacity (Przybylowski et al., 2021).

As a result of the effects of the COVID-19 pandemic, hence, obtaining information on on-board passenger comfort is now no longer just a desirable feature for operators and passengers (Zhang et al., 2017), but actually a necessary-one that can provide additional confidence to travellers and can, consequently, make a direct positive contribution to the economic viability and sustainability of public transport services (Transport Focus, 2021; Gkiotsalitis and Cats, 2021). Up until recently, due to the absence of the relevant enabling technology, the only way of obtaining the relevant data was through the conduct of exhaustive manual passenger counts.

* Corresponding author.
E-mail address: i.kaparias@southampton.ac.uk (I. Kaparias).

https://doi.org/10.1016/j.trc.2022.103963
Received 23 September 2021; Received in revised form 20 November 2022; Accepted 21 November 2022
Available online 1 December 2022
0968-090X/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

These would be carried out infrequently and would provide data of questionable accuracy, as they would only give a snapshot of the condition at the time of the survey. As a result, the key issue of estimating on-board comfort in real-time, which has now risen to prominence due to the COVID-19 effects, has received only little attention in the literature.

Nevertheless, technological advances in data acquisition, transmission and storage have accelerated the development and implementation of automated data collection systems (Koutsopoulos et al., 2017; Koutsopoulos et al., 2019), with Automatic Passenger Counting (APC) being a prominent example. APC systems make use of a standard system architecture of sensors on-board the public transport vehicle, thus providing much more accurate estimates of the vehicle on-board loads. As such, APC allows operators to gain an insight into the number of passengers boarding and alighting at each station, and therefore helps them in long-term planning, but can also, more crucially, assist them in real-time operations. APC systems are therefore increasingly being installed on public transport vehicles in various cities around the world. However, the cost of each APC installation and maintenance is high, and as a result only a small subset of the vehicles can usually be equipped. The common practice of public transport operators in France, for example, is that roughly 10–20 % of the total fleet is equipped, especially when legacy vehicles are considered. This means that only partial knowledge on the loadings can be obtained, which makes real-time loading predictions a particularly complex problem.

The aim of the present study is, hence, to propose an estimation method of on-board passenger loads, based on non-exhaustive (incomplete) APC datasets, in conjunction with Automatic Vehicle Locating (AVL) data, which is an Intelligent Transport Systems (ITS) feature that is installed almost as standard on public transport systems worldwide (Koutsopoulos et al., 2017; Koutsopoulos et al., 2019). The estimation is to be achieved per vehicle run, line stop, and time interval of a typical working day, in order to be suitable for use as input by operators in planning and management processes, particularly as concerns the provision of real-time crowding information to passengers in the post-COVID-19 era. The proposed method is then tested and validated using actual APC and AVL datasets from the tramway network of the French city of Nantes, covering a period of three months.

The rest of the paper is structured as follows. Section 2 reviews the background of the study, as reported in the available scientific literature in the related fields of passenger comfort measurement and quantification and public transport passenger loading forecasting and estimation methods. Section 3, then, presents the new on-board passenger load estimation method, including the formal problem definition, models for boarding and alighting passengers, and the resulting Kalman filter based approach used for the estimation. Section 4 provides a description of the study area and datasets, defines the evaluation metrics and processes, and reports and discusses the results obtained. Section 5 presents the numerical results produced with the proposed estimation method, considering various performance metrics. Section 6, finally, concludes the paper and identifies areas of further work.

2. Background

In order to establish the background of the present study, the relevant scientific literature is reviewed. This includes the topics of passenger comfort measurement and quantification, and passenger loading forecasting and estimation methods. These, then, lead to the identification of the research gap that the study addresses.

2.1. Passenger comfort measurement and quantification

With the increase of public transport patronage in the years prior to the COVID-19 pandemic, passenger comfort had already become a major issue in scheduling and operating public transport services, as it was seen as being related to significant welfare costs (Haywood et al., 2017). Comfort evaluation is, actually, a multi-criteria assessment problem, as defined by European Standard EN13816 (European Committee for Standardization, 2002). According to Mohammadi et al. (2020), the comfort level on-board public transport vehicles can be broken down to five critical factors: thermal, vibration, noise, lighting and air quality. Furthermore, the level of comfort can vary with respect to the volume of on-board passengers, where in-vehicle comfort and crowding have a significant impact on passenger (and, hence, customer) satisfaction (Cox et al., 2006), which is subjectively evaluated.

Consequently, the measurement of in-vehicle crowding can be performed using both subjective (perception-based) and objective (actually measured) metrics (Turner et al., 2005), and both have advantages and drawbacks. Subjective metrics, for instance, give a much more accurate idea of how passengers really perceive and rate their on-board experience. However, they are typically backed by very limited empirical evidence, and are also heavily influenced by external factors, such as geographical and cultural differences, which makes them difficult to assess, use, and generalise from (Li and Hensher, 2013). Objective metrics, on the other hand, may be much more difficult to relate to real passenger perceptions, but they can be much more easily measured and used as a common standard of what would constitute good or bad on-board comfort by most passengers. For example, Tirachini et al. (2012) evaluated a number of objective metrics, such as the density of standing passengers and the proportion of seats occupied on-board, and found that they represented a good approximation of what passengers actually experienced.

Looking at some examples of past studies on the topic, Kroes et al. (2013) conducted qualitative and quantitative surveys in order to quantify and measure in-vehicle comfort in the Paris metro. The study provided a typology of passengers with respect to their attitude towards travel time and comfort, which was obtained from stated preference survey models exploring the willingness to wait for a next, less crowded, train in relation to relative crowding levels. On the other hand, Haywood and Koning (2013) looked at the relationship between in and vehicle comfort and seating availability, and by carrying out surveys with public transport users in Paris, found that passenger inconvenience increased with decreasing in-vehicle comfort, and that there was a non-linear trade-off rate between "comfortable" and "uncomfortable" travel time. They also concluded that passengers are less keen on trading-off travel time for greater comfort in the morning peak hour, likely due to the constraints of morning commuting trips (e.g. punctual arrival at the workplace). Similarly, Batarce et al. (2015) evaluated in-vehicle comfort on the basis of mixed stated and revealed preference data in Santiago,

Chile, using discrete choice models, and found a twofold increase of the marginal disutility between a "low" density of 1 passenger/m^2 and a "higher" density of 6 passengers/m^2, linearly related with the travel time. Tirachini et al. (2017) built on that research to identify relationships between the crowding level, perceived comfort and security. This body of research motivates the definition of levels of in-vehicle comfort. For example, the US Transit Capacity and Quality of Service Manual (TCQSM) defines specific levels of on-board crowding from A to F, where the latter represents "crushing" loading levels (i.e., more than 5 standing passengers/m^2) (US Transportation Research Board, 2013).

On-board comfort, naturally, has direct implications on passenger demand and public transport operations. De Palma et al. (2015) discussed the necessity of distinguishing seating and standing as two different states of comfort and provided an analytical expression of the discomfort that can be employed in order to derive optimal timetables and tariffs. Several demand models followed the distinction of the two passenger states and provided specific algorithms to address the difference between them. For instance, Leurent and co-authors (Leurent and Liu, 2009; Leurent et al., 2013) built on previous models to formulate an integrated framework of transit assignment that considers comfort-related factors, such as train line capacity, vehicle passenger capacity and in-vehicle comfort. Trozzi et al. (2013) also provided a dynamic user equilibrium model for bus networks, considering capacity constraints at the vehicle level.

In light of COVID-19 and the associated changes in passenger habits and comfort thresholds, some of the assumptions behind several of these studies will likely need to be revisited. The principles, however, remain the same and highlight the need for a reliable way of estimating crowding levels on-board public transport vehicles.

2.2. Passenger loading forecasting and estimation methods

Several decades' worth of research have extensively explored the topic of travel demand estimation and forecasting. A considerable body of literature has focused on vehicle traffic; this is comprehensively appraised by Vlahogianni and co-authors (Vlahogianni et al., 2004; Vlahogianni et al., 2014), with methods typically being categorised into parametric and non-parametric ones. More recent research has attempted to transfer several of these methods onto the public transport domain in order to estimate or forecast passenger demand in the short-term. Examples of approaches adopted in this respect include autoregressive integrated moving average (ARIMA) and generalised autoregressive conditional heteroscedasticity (GARCH) (Ding et al., 2018), neural networks (Tsai et al., 2009; Jia et al., 2019); Kalman filtering (Guo et al., 2014; Gong et al., 2014), random forests (Cheng et al., 2019), and deep belief networks (Bai et al., 2017). A related problem that has also received considerable attention has been the estimation and prediction of Origin-Destination (OD) flows and matrices for public transport, usually on the basis of passenger counts or Automatic Fare Collection (AFC) systems. Methods adopted include optimisation (Gur and Ben-Shabat, 1997; Liu et al., 2021), elasticity (van Oort et al., 2015), trip chaining (Wang et al., 2011; Li et al., 2011), Iterative Proportional Fitting (IPF) (Ji et al., 2015), clustering (Huang et al., 2020), Bayesian inference (Sun et al., 2021), as well as data fusion and Kalman filtering (Tao and Tang, 2019). Some research has also explored the nature and patterns of prediction and forecasting errors and has created inferential statistics models aiming to address them (Jung and Casello, 2020).

The problem of estimating real-time on-board passenger loading, and consequently also passenger comfort, has received much less attention in the literature, however, primarily due to the lack of reliable data sources to date. Some research used manual passenger counts, such as for example, He et al. (2018), who developed a scheme employing Monte Carlo simulation, neural networks and Markov chains in order to more efficiently control bus air-conditioning systems in Beijing on the basis of the anticipated loading. Other research attempted to estimate on-board occupancy using WiFi, but with rather limited success, mostly due to the inability of probes to exclude WiFi-enabled devices outside the vehicle and to count passengers without WiFi-enabled devices on-board (Mikkelsen et al., 2016; Oransirikul et al., 2014).

More successful attempts have been carried out using a combination of AFC and AVL data. For instance, the method by Zhang et al. (2017) first estimated the on-board load of buses on the basis of trip chaining analysis and a probability model, and then predicted it using an extended Kalman filter, with promising results in terms of prediction accuracy when applied to the bus network of the city of Shenzhen. The approach of Noursalehi (2017), on the other hand, made use of random forests and gradient boosting for predicting passenger arrivals and their destinations on the London Underground network, along with an online simulation performing transit assignment. Random forests and gradient boosting, along with some other supervised learning methods (namely neural networks and k-Nearest-Neighbours), were also compared in the study by Heydenrijk-Ottens et al. (2018) for the prediction of both long- and short-term on-board loading of trams in the Hague, with, again, promising results.

Nevertheless, the main disadvantage of AFC is that in rail systems passengers are usually required to scan their smartcard at station entries and exits rather than on-board the vehicle, while in bus and tram systems they are usually only required to scan it when they board and not when they alight. As a result, a considerable amount of inference is needed in order to estimate on-board loads, which can make the process overly complex and can compromise accuracy. Sun et al. (2021) used vehicle dwell measured through AVL data to relate passenger flows to passenger activity and then formulate a Bayesian inference model to predict boarding and alighting flows, as well as passenger loads. APC systems can also make a difference, and several studies have made use of them recently. For example, Khomchuk et al. (2018) used a Bayesian estimation approach to predict train loads on the basis of real-time APC and historical data, which they validated on a simulated network, while Pasini et al. (2020) and Hu et al. (2020) both used neural networks to predict train loads in suburban Paris and the San Francisco Bay Area respectively based on temporal features and recent (same-day) previous measurements. Jung and Casello (2020) used AVL and APC data to examine transit ridership errors. Pasini et al. (2019) additionally experimented with time-series modelling and machine learning methods (specifically random forests and gradient boosting trees) and found that they were able to adequately consider the temporal irregularity of train services. Wang et al. (2021), on the other hand, developed a two-stage prediction process of bus passenger on-board loads, whereby an initial short-term prediction is effectuated using

an adaptive Kalman filter, and a further prediction is made using support vector regression; the performance of the method was then evaluated on the bus network of the city of Suzhou. Finally, Jenelius (2020) used a number of methods (stepwise regression, lasso regression and boosted tree ensembles) in order to predict real-time car-specific on-board crowding on the Stockholm metro network on the basis of APC (on-board passenger counts estimated through weight measurements of the train cars) and AVL data and found that when considering real-time data, the prediction accuracy improved.

APC systems, however, have two main limitations. The first limitation is that, due to the dynamic nature of the phenomenon observed (passengers entering and exiting buses, trams or trains), many of the enabling technologies (such as weight/pressure, optical or radar sensors) are unable to deliver high precision, which means that APC systems are often prone to downtime as well as measurement errors. The second limitation is that, as mentioned already, due to their high cost, APC systems would typically have very low penetration rates in a public transport fleet, and as such, they would only be able to deliver partial information. Several of the studies carried out so far have sufficiently addressed the first limitation (malfunctions), but have generally not dealt with the second one (the partial availability), having usually explicitly or implicitly assumed complete data availability. A notable exception to this has been the work of Jenelius (2019), who, extending their previous work (Jenelius, 2020), used their developed lasso regression approach to predict real-time on-board crowding on buses in Stockholm, taking into account the fact that only 20 % of the vehicles were equipped. The results suggested that run-specific load prediction improved as the target run approached the departure time from the station. While this approach is capable of delivering sufficiently accurate estimates, however, it requires a fairly extensive training and calibration phase every time it is applied on a new case study.

2.3. Summary

Consequently, from the review of the available literature, two research gaps can be identified. The first one is that, despite real-time and short-term passenger comfort information having been identified as an important factor of passenger mode choice, particularly post-COVID-19, and even though several passenger comfort quantification models have been developed, a link with on-board loading estimation has not been made to date. The second is that the majority of studies having attempted to estimate or predict on-board loading on the basis of APC have assumed access to complete datasets, which, however, is unrealistic in practice. Therefore, the present study addresses these gaps by developing a Kalman filter based on-board passenger comfort estimation method. An advantage of the approach is that it is largely off-the-shelf: as opposed to data-driven methods, it does not require a substantial amount of preparatory work to be carried out (such as data collection, data processing, model fitting, etc.), and is capable of producing estimates as soon as the first measurement point becomes available and of subsequently improving the accuracy of these estimates as more data become available. The approach is described in the next section.

3. On-board passenger load estimation methodology

This section presents the new estimation method of real-time in-vehicle comfort, as expressed by on-board passenger numbers, proposed by the present study. The overall estimation framework is described first, followed by the mathematical notation used and an outline of the modelling assumptions and conventions adopted. Dynamic models for representing boarding and alighting passengers are, then, formulated, and the observability of the proposed systems is analysed and assessed. The section is, then, concluded with the formulation of the proposed Kalman filter based estimation method.

3.1. Overall estimation framework

The aim of the proposed framework is the estimation the on-board comfort. “Estimation” here refers to the computation of the on-board comfort level at the current station and time. This is different to “prediction”, which refers to the establishment of the on-board comfort level at a later station and/or at a future time point, and which lies beyond the scope of the present study.

The proposed estimation method aims at taking advantage of the available measurements in terms of vehicle location and passengers. In particular, measurements originate from:

- AVL systems, installed in all vehicles, providing the position of a vehicle in real-time (including the time when a vehicle stops at each station); and
- APC systems, installed on a limited number of vehicles, providing the number of passengers on-board, as well as the numbers of boarding and alighting passengers at each station.

This results in a situation where full passenger information is available for some vehicles, but no passenger information is available for the remaining vehicles, which should therefore be estimated. However, this setting is not desirable for developing a vehicle-based estimation method, i.e., a method that directly processes vehicle-based measurements to calculate vehicle-based estimates, since no meaningful relation can be assumed between the passenger load of a vehicle and that of preceding (or subsequent) vehicles; this does not allow relating the available APC measurements to the quantities that are to be estimated.

On the other hand, a more reasonable way is to perform vehicle-based estimation by first estimating station-based quantities and then derive the vehicle-based quantities that are of interest. In fact, such “indirect” estimation is preferable, since, by casting the problem into an estimation problem of station-based quantities, partial passenger loading information is available for every station, i.e., provided when a vehicle with APC is at the station. Moreover, this allows formulating analytical (data-driven) models for station-

based passenger dynamics, which have a clear physical meaning and are based on reasonable assumptions, resulting in a more rigorous estimation problem. For instance, it is reasonable to assume that the number of passengers arriving at a station does not exhibit strong fluctuations at a given time of the day and, except in exceptional circumstances, is not affected by a variation in the public transport schedule, such as, for example, a minor train delay. On the other hand, an unexpected change of a vehicle headway may strongly affect the number of passengers boarding a vehicle and will likely affect also all successive runs.

There are multiple ways of modelling the passenger arrival, boarding, and alighting processes at a station. The very nature of the problem can result in some complex models characterised by several parameters, whose calibration will likely require the availability of vast quantities of data (e.g., Gur and Ben-Shabat, 1997; Liu et al., 2021; van Oort et al., 2015; Wang et al., 2011; Li et al., 2011; Ji et al., 2015; Huang et al., 2020; Sun et al., 2021; Tao and Tang, 2019). Here, simplified models are developed and employed. These are characterised by linear dynamics, which allow to represent the boarding and alighting processes at each station according to simplified, yet reasonable assumptions. These are, then, complemented by linear measurement models, which allow the incorporation of APC measurements, reformulated as station-based quantities, as well as historical information, obtained, for example, by pre-processing AVL and APC data from preceding days. The resulting models are therefore capable of assimilating real-time data, as well as historical data, resulting in a data fusion approach, which can be tailored to the data availability in order to achieve the best possible estimation performance.

Based on the developed models, the estimation is performed by employing a Kalman filter (KF) (Kalman and Bucy, 1961; Anderson and Moore, 1979), which is an effective methodology for state estimation of linear systems in the presence of limited and/or noisy measurements. The KF is an optimal state estimator applied to a dynamic system that involves random noise and includes a limited amount of noisy real-time measurements. In particular, the KF and its variants have been successfully applied in several domains, including transport (see, e.g., Szeto and Gazis, 1972; Wang and Papageorgiou, 2005; Bekiaris-Liberis et al., 2016; Roncoli et al., 2016; Bekiaris-Liberis et al., 2016; Antoniou et al., 2010; Achar et al., 2020).

To summarise, the proposed approach consists of the following basic components:

1. a station-based data-driven model for boarding passengers;
2. a station-based data-driven model for alighting passengers, formulated in terms of alighting rates;
3. the utilisation of vehicle position information and vehicle-based passenger measurements, where the latter are provided by a limited amount of vehicles equipped with APC systems;
4. the use of a KF for the real-time estimation of station-based boarding passengers and station-based alighting rates; and
5. a conservation-of-passengers equation for calculating the vehicle-based passenger load for each operating vehicle.

The different components are described in detail in the next sub-sections.

3.2. Problem notation, conventions, and assumptions

A public transport network is modelled by a set of stations I and a set of lines L, whereby an individual station is indexed by $i \in I$ and a specific line is indexed by $l \in L$. A single run of a public transport vehicle (train, tram or bus) along a certain line l is denoted $j \in J_l$, where J_l is the set of all runs along line l over an observation period, which is assumed being one full operational day. Here, dynamic models are considered, which are defined in the discrete-time domain, introducing a step size T (e.g. of the order of 30–120 s), where time is indexed by k, such that actual time $t = kT$.

The following variables are defined:

$p_j(k)$ number of passengers on-board run j at time.k

$b_j^{\mathrm{r}}(k)$ number of passengers boarding run j during.$(k-1,k]$

$a_j^{\mathrm{r}}(k)$ number of passengers alighting from run j during.$(k-1,k]$

$\gamma_j^{\mathrm{r}}(k)$ alighting rate of run j during.$(k-1,k]$

$w_{i,l}(k)$ number of passengers on the platform at station i waiting to board a vehicle of line l at time.k

$e_{i,l}(k)$ number of passengers entering station i platform to board a vehicle of line l during.$(k-1,k]$

$b_{i,l}^{\mathrm{s}}(k)$ number of passengers boarding a vehicle of line l at station i during.$(k-1,k]$

$a_{i,l}^{\mathrm{s}}(k)$ number of passengers alighting from a vehicle of line l at station i during.$(k-1,k]$

$\gamma_{i,l}^{\mathrm{s}}(k)$ alighting rate of vehicles of line l at station i during.$(k-1,k]$

$\eta_{i,j}^{\mathrm{r}}(k)$ binary variable indicating if run j is at (or departs from) station i during.$(k-1,k]$

$\eta_{i,l}^{\mathrm{s}}(k)$ binary variable indicating if a line l vehicle is at (or departs from) station i during.$(k-1,k]$

β_j^{r} binary variable indicating if run j is equipped with APC providing passenger load information.

$\beta_{i,l}^{\mathrm{s}}(k)$ binary variable indicating if a run of line l departing from station i during $(k-1,k]$ is equipped with APC, providing passenger load information.

φ binary variable indicating if historical data is used for estimation.

Also, for any variable ω, its measured value is denoted $\overline{\omega}$, its value calculated on the basis of "historical" observations is denoted $\widetilde{\omega}$, and its estimated value on the basis of the proposed method is denoted $\widehat{\omega}$.

The objective of the proposed method is to estimate the number of on-board passengers $p_j(k)$ for all runs over a certain period, by employing combined AVL and APC information, which is available from both historical and real-time data. It is assumed that the

estimation algorithm runs on a daily basis, considering an "operational" day, which typically starts in the morning of a calendar day (usually at 4 or 5 AM) and finishes in the early hours of the next calendar day (usually at 1 or 2 AM). Therefore, historical data comprise any data originating from previous "operational" days (which have been appropriately aggregated and pre-processed – an example of such pre-processing is documented in Section 4.3), while real-time data are received and processed whenever available, assuming no communication delays.

In developing the proposed estimation method, the following measurements are assumed to be available at any time k :

- Real-time AVL information for all runs and at all stations, providing $\eta^{r}_{i,j}(k), \forall i \in I, j \in J_l, l \in L$.
- Real-time APC data for a limited number of runs $\overline{J}_l \subset J_l$, providing $\overline{b}^{r}_{j}(k)$, $\overline{a}^{r}_{j}(k)$, and $\overline{p}_j(k)$, for $j \in \overline{J}_l$, $\forall l \in L$. This allows to assign $\beta^{r}_{j} = 1$ if $j \in \overline{J}_l$ and $\beta^{r}_{j} = 0$ otherwise.
- Historical information obtained by processing AVL and APC data available for the previous days, providing $\widetilde{e}_{i,l}(k)$ and $\widetilde{\gamma}^{s}_{i,l}(k), \forall i \in I$, $l \in L$.

Before proceeding to formulate station-based models, a correspondence between station-based quantities and vehicle-based quantities is formulated by first introducing two assumptions, which are, in general, trivially satisfied for public transport networks, considering a reasonably sized time-step T (e.g., of the order of 30 s to 2 min), depending on the resolution of the data and the public transport mode considered. For example, tram systems tend to exhibit longer headways and could accommodate a larger time step compared to urban bus systems that would require a shorter step size. In the test case provided in the following sections, the time-step is set 60 s to match the resolution of the data used. Specifically:

Assumption 1. *There is only one run j operating on line l that departs from station i during time interval* $(k-1, k]$, *i.e.*

$$\sum_{j \in J_l} \eta^{r}_{i,j}(k) = \eta^{s}_{i,l}(k), \forall i \in I, k. \tag{1}$$

Assumption 2. *A run j can depart from only one station during a time interval* $(k-1, k]$, *i.e.*

$$\sum_{i \in I} \eta^{r}_{i,j}(k) = 1, \forall j \in J_l, l \in L, k. \tag{2}$$

These assumptions allow introducing the following relations for boarding and alighting passengers:

$$b^{r}_{j}(k) = \sum_{i \in I} \eta^{r}_{i,j}(k) b^{s}_{i,l}(k), \forall l \in L, j \in J_l, k \tag{3}$$

$$a^{r}_{j}(k) = \sum_{i \in I} \eta^{r}_{i,j}(k) a^{s}_{i,l}(k), \forall l \in L, j \in J_l, k. \tag{4}$$

These imply that by estimating station-based quantities $b^{s}_{i,l}(k)$ and $a^{s}_{i,l}(k)$, vehicle-based quantities $b^{r}_{j}(k)$ and $a^{r}_{j}(k)$ can then be directly calculated. Hence, in the following sub-sections, models for estimating the former quantities are presented.

3.3. Dynamic model for boarding passengers

In order to estimate the number of passengers boarding a vehicle of line l at station i, $b^{s}_{i,l}(k)$, a dynamic model is introduced for the number of passengers on the platform at a stop waiting to board a run of a specific line. This evolves according to the following dynamics:

$$w_{i,l}(k+1) = w_{i,l}(k) + e_{i,l}(k) - b^{s}_{i,l}(k). \tag{5}$$

The following assumption is, then, introduced:

Assumption 3. *At the time that any vehicle operating on line l departs from station i, all passengers waiting on the platform to travel on line l at time k will board the vehicle during* $(k-1, k]$, *i.e.*,

$$b^{s}_{i,l}(k) = \eta^{s}_{i,l}(k) \cdot w_{i,l}(k) + \xi^{b}_{i,l}(k), \tag{6}$$

where $\xi^{b}_{i,l}(k)$ is an unknown modelling error, which can be, for example, described by zero-mean Gaussian noise. Describing random variables as zero-mean Gaussian is a typical approach in filtering design, as this allows specifying such stochastic process solely by its mean and variance, which, despite not matching exactly the process modelled, are deemed sufficient statistics for filtering purposes. In this case, a KF is employed, which has been rigorously proven optimal under the assumptions of a linear model and Gaussian noise. Still, it has been shown that one can successfully use KF even when the noise is not Gaussian (as almost always the case in real life), and that makes KF the best linear filter (Simon, 2006).

It should be noted that Assumption 3 is typically satisfied in public transport networks, where there are no passengers left behind.

However, situations of extreme passenger congestion can occur in practice in some public transport networks during peak times, and in such cases the assumption does not hold. This, however, is not a limitation of the proposed model, but rather an inherent limitation of APC as a measurement technology. This is because APC is capable of capturing only the passengers that board a vehicle but is unable to provide any information on the actual demand of passengers (and/or any left-behind passengers).

Substituting (6) into (5) leads to:

$$w_{i,l}(k+1) = \left[1 - \eta^{s}_{i,l}(k)\right] \cdot w_{i,l}(k) + e_{i,l}(k) + \xi^{b}_{i,l}(k). \tag{7}$$

Since there is no available information on the number of passengers entering the platforms to board vehicles, $e_{i,l}(k)$ is treated as constant (or, effectively, slowly varying), being characterised by random walk dynamics,[1] i.e.

$$e_{i,l}(k+1) = e_{i,l}(k) + \xi^{e}_{i,l}(k), \tag{8}$$

where $\xi^{e}_{i,l}(k)$ is, for example, zero-mean Gaussian noise. Although this may seem a crude approach, such simplified dynamic model is widely used for model-based estimation in the absence of a descriptive dynamic model (e.g., Wang and Papageorgiou, 2005; Bekiaris-Liberis et al., 2016).

The overall (deterministic part of) system (7)-(8) is next written in a compact state-space form by defining the state vector of the system as

$$x^{b}_{i,l} = \begin{bmatrix} w_{i,l} \\ e_{i,l} \end{bmatrix}, \tag{9}$$

whose dynamics evolve according to

$$x^{b}_{i,l}(k+1) = A^{b}(k) \cdot x^{b}_{i,l}(k), \tag{10}$$

where

$$A^{b}(k) = A^{b}\left(\eta^{s}_{i,l}(k)\right) = \begin{bmatrix} 1 - \eta^{s}_{i,l}(k) & 1 \\ 0 & 1 \end{bmatrix}. \tag{11}$$

System (10)-(11) is a linear-parameter-varying (LPV) system, where parameter $\eta^{s}_{i,l}(k)$ is assumed to be known (measured), as stated in Section 3.2.

Available real-time measurements for system (10)-(11) are obtained from APC measurements, which are, however, available only when a run equipped with APC is leaving the station, i.e., when $\beta^{s}_{i,l}(k) = 1$, where $\beta^{s}_{i,l}(k)$ is calculated from measured quantities as

$$\beta^{s}_{i,l}(k) = \sum_{j \in J_l} \beta^{r}_{j} \cdot \eta_{i,j}(k). \tag{12}$$

Following the rationale of Assumption 3 and (12), when an APC equipped vehicle of line l is at station i, we can treat the available measurement for $b^{s}_{i,l}(k)$ as a (noisy) measurement for $w_{i,l}(k)$; this is formulated as

$$z^{w}_{i,l}(k) = \beta^{s}_{i,l}(k) \cdot w_{i,l}(k) + \psi^{w}_{i,l}(k), \tag{13}$$

where $\psi^{w}_{i,l}(k)$ is a measurement error in the form of zero-mean Gaussian noise.

Nevertheless, as real-time data are available only at specific times, i.e., when a vehicle with APC is at or departs from a station ($\beta^{s}_{i,l}(k) = 1$), there could be long periods for which no measurements are available. This may cause issues due to, for example, daily recurrent fluctuations of passenger arrivals, demand peaks, etc., which, if not "observed" via a measurement, may cause a deterioration of the estimation performance. In order to overcome this issue, it is proposed to employ also historical data to feed the estimator when real-time information is not available. In particular, availability of historical data is assumed in terms of the number of passengers entering the platform of a station to board a specific line $\widetilde{e}_{i,l}(k)$, which can be obtained by processing AVL and APC data from previous days. The resulting measurement model reads:

$$z^{e}_{i,l}(k) = \varphi\left(1 - \beta^{r}_{i,l}(k)\right) \cdot \widetilde{e}_{i,l}(k) + \psi^{e}_{i,l}(k), \tag{14}$$

where $\psi^{e}_{i,l}(k)$ is the measurement error associated with historical data in the form of a zero-mean Gaussian noise and φ is a binary parameter indicating whether historical data are used ($\varphi = 1$) or not ($\varphi = 0$).

To summarise, system (10)-(11) is complemented by associating an output vector $z^{b}_{i,l}$, described by the following (deterministic part of the) measurement model:

[1] A *random walk* is an approach for modelling a stochastic process composed of a series of random variables through time. A Gaussian random walk is used in this study, in which the time series data are assumed to be generated based on a normal distribution.

$$z^b_{i,l}(k) = \begin{bmatrix} z^w_{i,l}(k) \\ z^e_{i,l}(k) \end{bmatrix} = C^b(k) \cdot x^b_{i,l}(k), \tag{15}$$

where $C^b(k)$ is obtained by known (measured) parameters as

$$C^b(k) = C^b\left(\beta^s_{i,l}(k), \varphi\right) = \begin{bmatrix} \beta^s_{i,l}(k) & 0 \\ 0 & \varphi\left(1 - \beta^s_{i,l}(k)\right) \end{bmatrix} \tag{16}$$

The noisy (measured) version of $z^b_{i,l}(k)$, which holds the passenger measurements that are available for estimation, either from real-time APC data or from historical data, is:

$$\overline{z}^b_{i,l}(k) = \begin{bmatrix} \overline{w}_{i,l}(k) \\ \widetilde{e}_{i,l}(k) \end{bmatrix}, \tag{17}$$

where $\overline{w}_{i,l}(k)$ is obtained from measured quantities available at any time step k as

$$\overline{w}_{i,l}(k) = \sum_{j \in J_l} \eta^r_{i,j}(k) \cdot \overline{b}^r_j(k). \tag{18}$$

3.4. Dynamic model for alighting passengers

A second model for estimating the number of passengers alighting at any station of a line is now formulated. In this case, instead of modelling directly the number of alighting passengers, a relationship between the number of passengers on-board a vehicle and the number of passengers alighting at a station is introduced, namely:

$$\gamma^r_j(k) = \frac{a^r_j(k)}{p_j(k)}, \forall l \in L, j \in J_l, k. \tag{19}$$

As previously stated, real-time estimation of vehicle-based variables is challenging, since any small disturbances in the schedules or passenger patterns may create an estimation bias that would be difficult to identify and correct. For this reason, the alighting rate is re-defined as a station-based variable, denoted by $\gamma^s_{i,l}(k)$, as (from Assumption 2):

$$\gamma^s_{i,l}(k) = \sum_{j \in J_l} \eta^r_{i,j}(k) \gamma^r_j(k), \forall i \in I, l \in L, k. \tag{20}$$

Variable $\gamma^s_{i,l}(k)$ represents the percentage of passengers on-board any vehicle operating on line l that alights at station i. Under the reasonable assumption that such value does not feature strong fluctuations in time (i.e. can be considered as slowly-varying), and in absence of a descriptive dynamic model, its dynamics is modelled as constant via a random walk, i.e.

$$\gamma^s_{i,l}(k+1) = \gamma^s_{i,l}(k) + \xi^\gamma_{i,l}(k), \tag{21}$$

where $\xi^\gamma_{i,l}(k)$ is, for example, zero-mean Gaussian noise. It should be noted that the deterministic part of system (21) is a linear time-invariant (LTI) system.

Real-time measurements for system (21) are again assumed to be available when an APC-equipped vehicle of line l is at station i; this results in the following measurement equation:

$$z_{i,l}^{\overline{\gamma}}(k) = \beta^s_{i,l}(k) \cdot \gamma^s_{i,l}(k) + \psi_{i,l}^{\overline{\gamma}}(k), \tag{22}$$

where $\psi_{i,l}^{\overline{\gamma}}(k)$ is the measurement error associated with real-time data in the form of a zero-mean Gaussian noise. Similarly as for the boarding passenger model, since real-time data are available only at specific times, i.e. when a vehicle with APC is at or departs from a station ($\beta^s_{i,l}(k) = 1$), the measurement data may be complemented by historical information that is fed to the estimator when there is no real-time information available. In this case, availability of historical data on the alighting rate $\widetilde{\gamma}^s_{i,l}(k)$ is assumed, which can be extracted by processing AVL and APC data from previous days. The resulting measurement model reads:

$$z_{i,l}^{\widetilde{\gamma}}(k) = \varphi\left(1 - \beta^s_{i,l}(k)\right) \cdot \widetilde{\gamma}^s_{i,l}(k) + \psi_{i,l}^{\widetilde{\gamma}}(k), \tag{23}$$

where $\psi_{i,l}^{\widetilde{\gamma}}(k)$ is the measurement error associated with historical data in the form of a zero-mean Gaussian noise. Therefore, to complete model (21), defined for passenger alighting rate, the output vector $z^\gamma_{i,l}$ is introduced, described by the following (deterministic part of the) measurement model:

$$z^\gamma_{i,l}(k) = C^\gamma(k) \cdot \gamma^s_{i,l}(k), \tag{24}$$

where $C^{\gamma}(k)$ is obtained from measured quantities available at any time step k as:

$$C^{\gamma}(k) = C^{\gamma}\left(\beta^{s}_{i,l}(k), \varphi\right) = \begin{bmatrix} \beta^{s}_{i,l}(k) \\ \varphi\left(1 - \beta^{s}_{i,l}(k)\right) \end{bmatrix} \tag{25}$$

The noisy (measured) version of $z^{\gamma}_{i,l}(k)$, which holds all the passenger measurements that are available for estimation, either from real-time APC data or from historical data, is:

$$\bar{z}^{\gamma}_{i,l}(k) = \begin{bmatrix} \bar{\gamma}^{s}_{i,l}(k) \\ \tilde{\gamma}^{s}_{i,l}(k) \end{bmatrix}, \tag{26}$$

where $\bar{\gamma}^{s}_{i,l}(k)$ is obtained from measured quantities available at any time step k as:

$$\bar{\gamma}^{s}_{i,l}(k) = \sum_{j \in J_l} \eta^{r}_{i,j}(k) \cdot \frac{\bar{a}^{r}_{j}(k)}{\bar{p}_{j}(k)}. \tag{27}$$

3.5. Observability of the proposed systems

Before proceeding to design estimators for boarding and alighting passengers, the observability of the systems formulated in the previous sections is investigated. In order to support readers that may not be familiar with the concept of observability, some physical implications of the formal definitions of observability are provided first (e.g., Antsaklis and Michel, 2006; Liu et al., 2013). In simple terms, the observability property of a system guarantees that the dynamic evolution of its internal states (i.e., the states that are not directly measured) may be determined (observed) by measuring only some specific states (or, more generally, some outputs of the system). In particular, while dealing with real-time state estimation, observability is a property that guarantees that the state of a system, such as the boarding passengers or alighting rates, can be reproduced, in real-time in an unbiased way from the available (partial) measurements by use of an estimator, such as a KF.

The observability of a system is usually studied employing certain algebraic conditions (see, e.g., Antsaklis and Michel, 2006), related in particular to the A and C matrices characterising the system. However, for time-varying or parameter-varying systems (as in the case of the models considered in this study), it may not be trivial to formally check and guarantee these conditions, since the parameters affect the system's matrices in real-time. For this reason, an alternative graph-theoretic approach is employed, which allows studying the observability property of a system by looking into its structure, defined by the zero and non-zero elements of the A and C matrices (see, e.g., Liu et al., 2013; Lin, 1974; Reissig et al., 2014). Moreover, the study of the structural observability properties of a system is useful in order to determine under which measurement configurations a system is actually observable.

It should be noted that structural observability is a necessary condition for observability, as it provides an intuitive way to the study of observability which, in practice, typically implies, indeed, system observability. However, the loss of observability of a structurally observable system may happen for some time intervals as a consequence of a combination of parameters that cause the elements of the A and C matrices to satisfy some specific conditions (e.g., Liu et al., 2013; Lin, 1974; Reissig et al., 2014). On the other hand, if no combinations of parameters guaranteeing (structural) observability exist, no estimator would be able to reconstruct the system state from the measured outputs. Thus, in practice, as it is also suggested from the estimation results in Section 5, structural observability implies a proper operation of an estimation scheme as the one presented here, even though the system may not always be formally completely observable at any time.

In order to study the structural observability for the proposed systems, the structure matrices $\mathscr{A}$ and $\mathscr{C}$ are introduced, representing the patterns of zero and non-zero elements of system matrices A and C, respectively. A useful representation of such patterns is via the construction of graphs $\mathscr{G}\left(\mathscr{A}^T, \mathscr{C}^T\right)$, which are shown in Fig. 1 for both the boarding passenger model, considering (10) and (15), and the alighting rate model, considering (21) and (24).

The following condition for structural observability is considered (as per, for example, Liu et al., 2013; Lin, 1974):

Condition 1: A linear system (A, C) is structurally observable if and only if: i) The graph $\mathscr{G}\left(\mathscr{A}^T, \mathscr{C}^T\right)$ contains no non-accessible vertex; and ii) the graph $\mathscr{G}\left(\mathscr{A}^T, \mathscr{C}^T\right)$ contains no dilation.

Considering the definition stated above, it can be established that both systems generally satisfy the conditions for structural observability, that is that there exist combinations of parameters that guarantee observability. This can be demonstrated by assuming $\beta^{s}_{i,l} = 1$ (or, more generally, non-zero) and observing that, for both systems, all vertices can be accessed, while no dilation exists in the graphs.

In addition, Condition 1 allows determining for which combination of parameter values the system is observable or when it may temporarily lose observability; this can be investigated by looking at the resulting graphs when some of the dashed edges are removed. In particular, the following claims are established:

1. If historical data are utilised for the boarding passenger model ($\varphi = 1$), when $\beta^{s}_{i,l}(k) = 0$ the system is temporally only partially observable due to the non-accessibility of vertex $w_{i,l}$, while vertex $e_{i,l}$ remains always accessible.

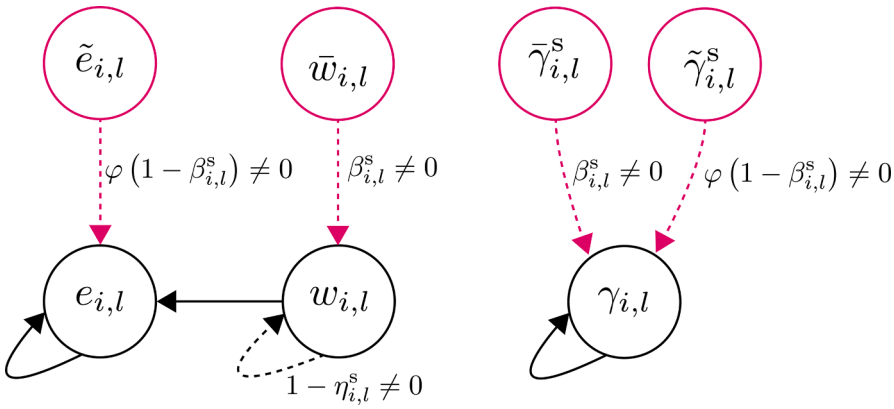

Fig. 1. The graphs $\mathscr{G}(\mathscr{A}^{T}, \mathscr{C}^{T})$ for patterns $\mathscr{A}$ and $\mathscr{C}$ that include matrices $\boldsymbol{A}$ and $\boldsymbol{C}$, respectively, of system (10), (15) (left), and of system (21), (24) (right). Black circles relate to the process models and red circles relate to the measurement models. Dashed lines indicate that the edge may exist, depending on the condition of parameters listed next to it (from which the time dependence is omitted).

2. If historical data are not utilised for the boarding passenger model ($\varphi = 0$), the system is temporally not observable when $\beta^{s}_{i,l}(k) = 0$ due to the non-accessibility of both vertices.
3. For both previous cases related to the boarding passenger model, observability is fully restored when $\beta^{s}_{i,l}(k) = 1$ (i.e., when a vehicle operating on line l equipped with APC stops at station i).
4. If historical data are utilised for the alighting rate model ($\varphi = 1$), the system is always observable, since, irrespectively of the value $\beta^{s}_{i,l}(k)$, one of the two dashed edges is present.
5. If historical data are not utilised for the alighting rate model ($\varphi = 0$), the system is temporally not observable when $\beta^{s}_{i,l}(k) = 0$, since no dashed lines are present.

Thus, in practice, as will also be shown by the estimation results in Section 5, apart from the cases described above, in which observability conditions are not met (partially, i.e., only for some states, or completely, i.e. for all states), the structural observability property holds and implies, as a general rule, the proper operation of an estimation scheme like the one proposed by the present study. In fact, since the cases in which observability is lost are only temporary occurrences, at the time when observability is restored the estimation capabilities of the proposed scheme are again guaranteed. Finally, it is noted that, if neither historical nor real-time data are available at any time, the system would be unobservable.

3.6. Estimation method

The KF algorithm that is employed to estimate boarding passengers and alighting rates using the models previously described is introduced here. The estimation equations for a KF are given by:

$$\hat{x}^{-}(k) = A(k-1)\hat{x}(k-1) \tag{28}$$

$$P^{-}(k) = A(k-1)P^{+}(k-1)A(k-1)^{\mathrm{T}} + Q \tag{29}$$

$$K(k) = P^{-}(k)C(k)^{\mathrm{T}}\left[R + C(k)P^{-}(k)C(k)^{\mathrm{T}}\right]^{-1} \tag{30}$$

$$\hat{x}(k) = \hat{x}^{-}(k) + K(k)[\bar{z}(k) - C(k)\hat{x}(k)] \tag{31}$$

$$P^{+}(k) = [I - K(k)C(k)]P^{-}(k), \tag{32}$$

where $\hat{x}^{-}$ and $\hat{x}$ denote, respectively, the a-priori (i.e. predicted) and a-posteriori (i.e. updated) estimates of variable (vector) x; $\bar{z}$ is a (noisy) measurement of x; A and C describe the state-transition and observation models of x; P^{-} and P^{+} are the a-priori (i.e. predicted) and a-posteriori (i.e. updated) estimated co-variance matrices; K is the optimal Kalman gain; and variables $Q = Q^{T} > 0$ and $R = R^{T} > 0$ are tuning parameters that represent the (ideally known) covariance matrices of the process and measurement noise, respectively.

Eq. (28) calculates the predicted (a priori) state estimate, i.e., the estimate of the system's state considering the previous (estimated) state and the system dynamics, whereas (29) calculates the predicted (a priori) covariance, i.e., a measure of the estimated uncertainty of the prediction of the system's state when employing only the system's dynamics. Eq. (30) calculates the optimal Kalman gain K, i.e., the gain that minimises the residual error in the minimum mean-square-error sense. Finally, Eq. (31) calculates the updated (a posteriori) state estimate, accounting for the correction due to the available measurements, while (32) calculates the updated (a posteriori) estimate covariance, i.e., a measure of the estimated uncertainty of the prediction of the system's state after measurements

are taken into account.

The algorithm is initialised as:

$$\widehat{x}(k_0) = \mu \tag{33}$$

$$P(k_0) = H, \tag{34}$$

where μ and $H = H^T > 0$ represent, in the ideal case where $\widehat{x}(k_0)$ is a Gaussian random variable, the mean and auto-covariance of $\widehat{x}(k_0)$ and $P(k_0)$, respectively.

In particular, two separate KFs are implemented: one for the estimation of boarding passengers and one for the estimation of alighting rates.

For estimating boarding passengers, an estimator for $x^b_{i,l}$ is designed considering process model (10)-(11), measurement model (15)-(16), and employing measurements (17); moreover, initial values $\mu = \begin{bmatrix} 0 & e^0_{i,l} \end{bmatrix}^{\mathrm{T}}$ and $H = I_{2\times 2}$ are considered. The estimator delivers estimates $\widehat{x}^b_{i,l}$, from which $\widehat{w}_{i,l}$ is to be extracted; the estimated boarding passengers can be derived from (3) and (6) as:

$$\widehat{b}^{\mathrm{r}}_j(k) = \sum_{i\in I} \eta^{\mathrm{r}}_{i,j}(k) \cdot \widehat{w}_{i,l}(k), \forall j \in J_l. \tag{35}$$

For estimating alighting passengers, an estimator for $\gamma^{\mathrm{s}}_{i,l}$ is designed, considering process model (21), measurement model (24)-(25) and employing measurements (26); initial values are set as $\mu = \gamma^{\mathrm{s},0}_{i,l}$ and $H = 1$. The estimator delivers estimates $\widehat{\gamma}^{\mathrm{s}}_{i,l}$, from which $\widehat{\gamma}^{\mathrm{r}}_j$ is calculated as:

$$\widehat{\gamma}^{\mathrm{r}}_j(k) = \sum_{i\in I} \eta^{\mathrm{r}}_{i,j}(k) \cdot \widehat{\gamma}^{\mathrm{s}}_{i,l}(k). \tag{36}$$

In both cases, it is possible that the KF delivers negative estimates at some steps, which are physically unrealistic. This is handled here in a heuristic manner, by bounding, at each step, the resulting estimates to be non-negative, and then using the bounded value at the next iteration. Even though some more complex methods exist to deal with this issue (Simon, 2010), testing them here led to virtually identical results.

Finally, in order to estimate the passengers on board of vehicle j, $\widehat{p}_j(k)$, a conservation-of-passengers equation is employed at each time step, of the form

$$\widehat{p}_j(k+1) = \widehat{p}_j(k) + \widehat{b}^{\mathrm{r}}_j(k) - \widehat{a}^{\mathrm{r}}_j(k), \tag{37}$$

where $\widehat{p}_j(0) = 0$, i.e. the vehicle is empty at the beginning of the service. Combining (37) with (19) results in:

$$\widehat{p}_j(k+1) = \left[1 - \widehat{\gamma}_j(k)\right]\widehat{p}_j(k) + \widehat{b}^{\mathrm{r}}_j(k), \tag{38}$$

which is calculated at each discrete time interval after estimates for $\widehat{b}^{\mathrm{r}}_j$ and $\widehat{\gamma}^{\mathrm{r}}_j$ are computed.

The overall estimation methodology is illustrated in Fig. 2.

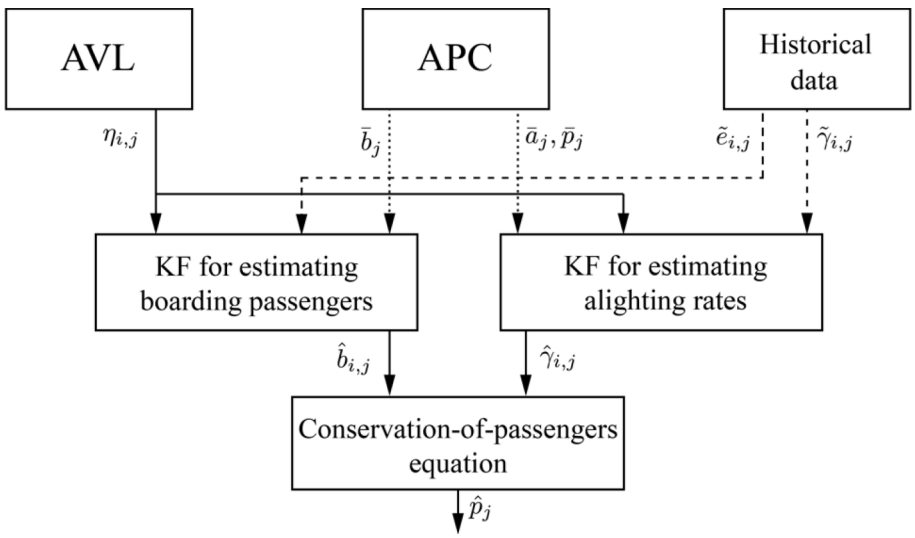

Fig. 2. The proposed estimation scheme.

4. Data acquisition and processing

The developed estimation method for real-time on-board passenger loads on the basis of AVL and APC data is applied on a real public transport network in this study, and this section sets out the core principles and methods used in that respect. The study area and dataset are introduced first and are followed by an outline of the data cleansing and processing tasks and by a description of the assimilation of the historical dataset used in the estimation. Finally, a brief description of the in-vehicle comfort measurement framework used is provided.

4.1. Study area and dataset

The present study focuses on the tramway system of the French city of Nantes. Nantes is located on the Loire River in Western France, close to the Atlantic coast. It is the sixth largest city of France, with a metropolitan population of 900,000. Its tramway network is operated by Semitan, and with its opening in 1985 Nantes became the first city to introduce a modern generation tramway, built from scratch. With its subsequent extensions, the network now consists of three tramway lines (numbered 1, 2 and 3) running on 44 km of track and serving a total of 83 stations, as well as a "Busway" Bus Rapid Transit (BRT) line (numbered 4).

The Nantes tramway is shown in Fig. 3. Line 1, shown in green, has a length of 18.4 km and serves 34 stations. It consists of two branches at each end (Beaujoire and Ranzay in the East, and François Mitterand and Jamet in the West) and a central trunk between the branches with 19 stations. Its frequency reaches 15 vehicles per hour during peak times, and it is the busiest line on the network (and with 120,000 passengers per day, it is also one of the busiest of the whole of France), serving several principal locations of the city, including the main railway station and the city's stadium. Line 2, shown in red, runs from Orvault in the North to Gare de Pont-Rousseau in the South, has a length of 11.7 km and serves 25 stations, including important educational (university) and health establishments. It has a frequency of 8 vehicles per hour during peak times, and its patronage approaches roughly 80,000 passengers per day. Lastly, Line 3, shown in blue, runs from Marcel Paul in the North to Neustrie in the South, has a length of 14.1 km and serves 34 stations. It has a similar operation with Line 2, with which it shares the track for seven stations (Commerce to Gare de Pont Rousseau) in the city centre. It serves several major commercial sites and is used by 75,000 passengers per day. The three lines run radially off the city centre but meet at Commerce. They are combined with Park and Ride (P + R) facilities on the outskirts, and also have major transfer points with the other public transport modes: the Busway (exclusive right-of-way line), the Chronobus (buses with limited segregated lines), the local buses and the regional coaches.

The tramway system is served by three types of rolling stock, irrespectively of the line: the Alstom TFS, the Bombardier Incentro, and the CAF Urbos. The Alstom TFS is a 39 m long vehicle with a capacity of 236 passengers (including 74 seats) which began operation in 1985. Each Alstom vehicle is composed of two high floor carriages with three-step accesses (of which one mobile step) and a lower floor carriage in the middle; access is provided by six double length doors and two simple doors per vehicle side. The Bombardier Incentro is 36 m long with a capacity of 252 passengers (including 72 seats) and started operating in 2000. It has an integral low floor and six double (1.30 m) doors per side. Finally, the CAF Urbos is the newest vehicle in the network, having started operations in 2012. It is 37 m long with a capacity of 249 passengers (including 68 seats) and has an integral low floor and six double doors per vehicle side.

The data used in this study have been collected from the Opthor and Ineo systems, used by the operator. Opthor is an APC system measuring the number of passengers boarding and alighting at each station, as well as the dwell time and other performance-related

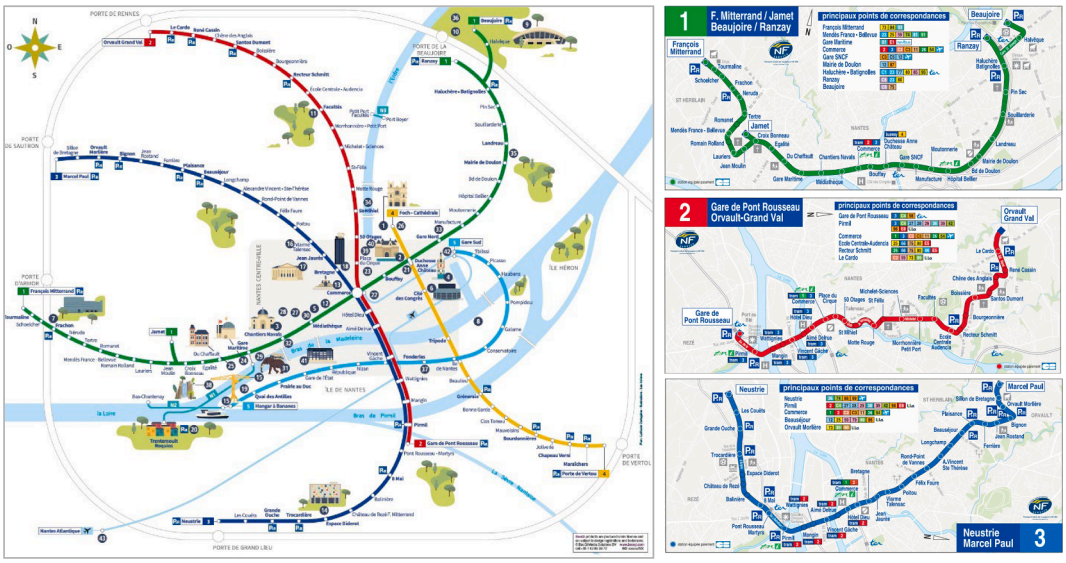

Fig. 3. The Nantes tramway network (Source. https://www.tan.fr).

measures. The system detects passengers using infrared sensors installed at each door of the vehicle with a 95 % measurement accuracy. Ineo, on the other hand, is an AVL system that enables comprehensive real-time tracking of each tramway vehicle, and can therefore deliver data on the actual arrival and departure time of each vehicle at each stop. All vehicles circulating in the network are equipped with Ineo; however, the operator has only installed Opthor on a limited number of vehicles (of Alstom TFS and Bombardier Incentro class only) and uses it to run frequent counts with it for various purposes.

4.2. Dataset processing and cleansing

The passenger data employed here were collected during an Opthor count conducted between 16 September and 20 December 2019 on Line 1 of the Nantes tramway network. As part of the count, the operator collected 52,350 valid entries from 42 weekdays and a total of 1946 runs. This sample corresponds to roughly 10 % of the total number of daily runs. Each data entry reports the total number of boarding and alighting passengers (as detected by the APC sensors), as well as the number of passengers on-board. The Opthor data are, then, complemented by the corresponding Ineo dataset for the same period. Fig. 4 illustrates the number of runs covered by AVL and APC data throughout the study period; days with no APC data availability or AVL malfunctions have been discarded.

A two-step process is performed to clean the AVL data and merge them with the APC relevant data. First, the AVL data are processed in order to remove erroneous data and fill missing data. This cleaning process involves correcting entries from inaccurate geo-localisation, namely:

- removal of double entries of a station for a single run;
- addition of missing stops, including terminal stops, to specific runs, and inference of travel times through distance-based linear interpolation between known stops; and
- removal of runs that do not provide service coherence.

Consequently, out of 20,360 commercial runs throughout the study period monitored by the AVL systems, 18,672 valid runs are kept.

The APC and the AVL systems are operated independently and use different data identifiers (station and run IDs, as well as timestamps), and as such an exact matching of APC runs with the "clean" AVL runs does not exist. Therefore, a run-matching process by approximation is employed in order to ensure as good a match as possible between the datasets, as the second step of the process. The method relies on a matching algorithm employed for each day and for each origin station throughout the day, which involves conducting a matching test for each AVL run with respect to selected "matching candidate" APC runs. According to the algorithm, a match between an AVL run and an APC run is established if the AVL and APC runs share the same stops, and:

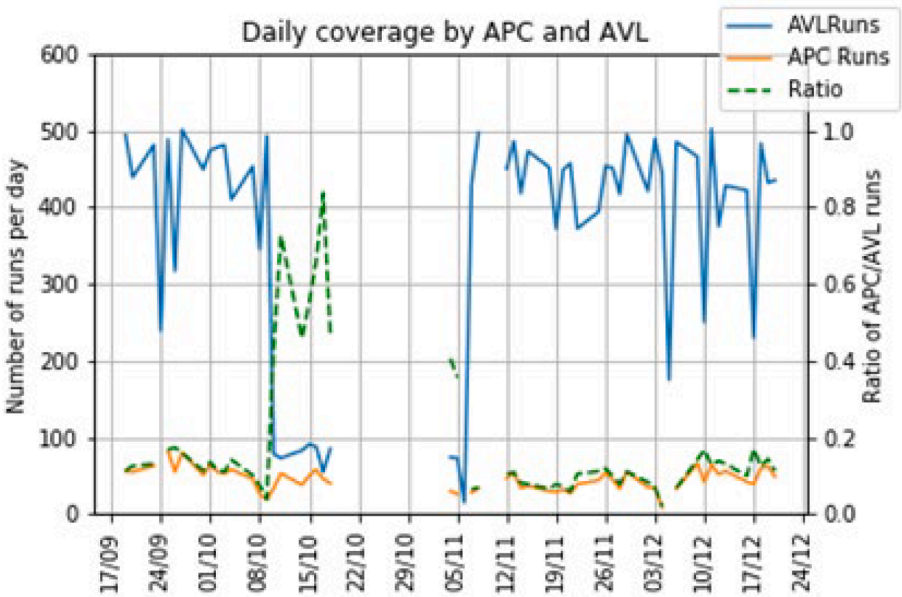

Fig. 4. Daily APC and AVL coverage during the study period.

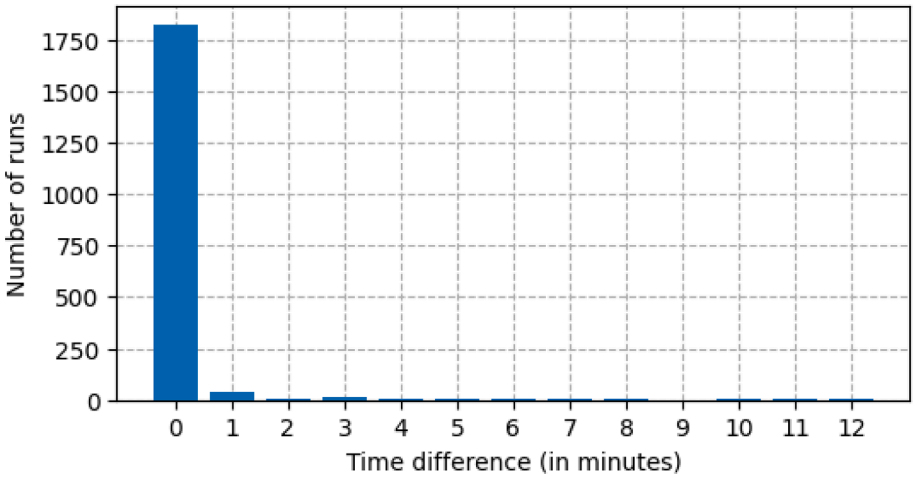

Fig. 5. AVL-APC time difference distribution for matched runs.

- the departure times of the AVL run and the APC run (either the theoretical/scheduled departure time or the actual departure time) are within 2 min of each other; or
- the departure times are within 2–15 min and the APC run is the one nearest to the AVL run.

If the difference between the departure times of the APC and AVL runs is greater than 15 min, the match is considered invalid and the relevant APC run is discarded.

The parameters used in the run-matching process are justified as follows. Firstly, the 2-min threshold value determining immediate run-matching is based on the usual headway between runs at peak-time, as per the present dataset and the authors' experience with other similar datasets. Then, the reason why both theoretical/scheduled and actual departure times from the AVL data need to be considered is the fact that the actual departure time may be inaccurate due to a vehicle dwelling at the origin station between runs without the AVL detectors on the ground distinguishing between its arrival (in the previous run) and departure (in the next run). Finally, the 15-min threshold is used in order to allow matching an APC run with the next recorded AVL run, if the latter is affected by lower geo-localisation precision accuracy – which can be a common source of error in AVL data.

Out of the 1946 initial APC runs from the entire study period, 29 (1,5%) are discarded by the run-matching process, leaving 1917 APC runs matched with an AVL run. Of the matched runs, 95 % have a time difference of up to 1 min, with only 1 % greater than 5 min, as illustrated in Fig. 5.

4.3. Assimilation of historical data

The methodology presented in Section 3 requires availability of historical data, which can be determined by processing AVL and APC data from days prior to the one for which estimation is carried out. In particular, historical data are required in order to compute $\widetilde{e}_{i,l}(k)$ and $\widetilde{\gamma}^{s}_{i,l}(k)$, i.e. the number of passengers entering the platform of station i to board a vehicle of line l and the alighting rate of vehicles of line l at station i, as well as $e^{0}_{i,l}$ and $\gamma^{s,0}_{i,l}$, i.e. the initial values for the estimates applied at the beginning of an operational day.

The historical data extracted from the APC counts may vary with respect to the day of the week. As suggested previously, historical data observations are not homogenously distributed throughout the days of the week. In addition, a number of days (specifically Tuesdays) have been discarded from the initial dataset due to technical issues and major incidents. Furthermore, some variance is observed at the within-day profile of the number of boardings per stop, run and hour (Fig. 6). For example, Friday has a higher afternoon peak and the service is extended at night, whereas Wednesday has a higher midday peak. These differences suggest only broad day-to-day regularity and no within-day regularity, which is consistent with the mobility patterns of French urban areas (De Solere, 2012).

Consequently, two strategies for calculating historical data are employed here: a) aggregating data from the same weekday; and b) aggregating all available days together. Regardless of the aggregation strategy, historical data are averaged and aggregated into a set of bins, each of a 30-min duration, that include all APC measurements available during such time. As the operational day starting at 4:00 AM and ending the next day at 3:59 AM is considered, a total of 48 30-min bins are used. The historical value $\widetilde{e}_{i,l}\left(\widetilde{k}\right)$, i.e. the historical value of the number of passengers per interval T entering the platform of station i to board a vehicle of line l during $(\widetilde{k}\widetilde{T}, \left(\widetilde{k}+1\right)\widetilde{T}-1]$, is then calculated as

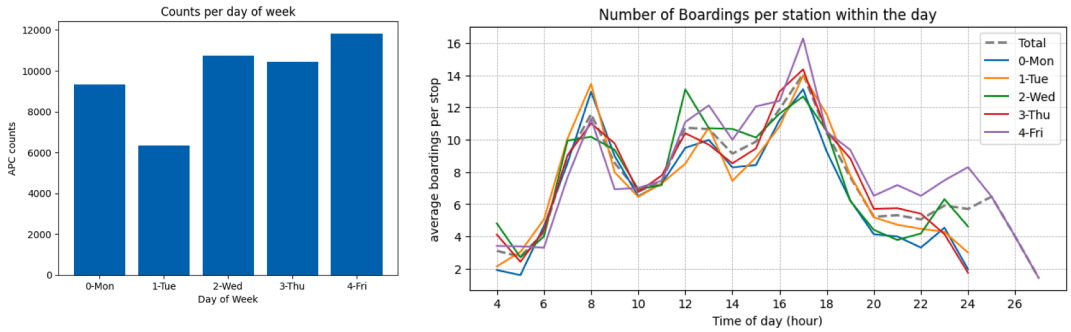

Fig. 6. Observations per day of the week (left) and average volume of boardings per stop and run per hour and per day of the week (right).

$$\widetilde{e}_{i,l}\left(\widetilde{k}\right) = \frac{\sum_{d\in D}\sum_{\widetilde{k=kT}}^{\widetilde{(k+1)}\widetilde{T}} - 1\frac{\overline{b}^{\mathrm{d}}_{i,l,d}(k)}{t^{\mathrm{d}}_{i,l,d}(k)}}{\sum_{d\in D}\sum_{\widetilde{k=kT}}^{\widetilde{(k+1)}\widetilde{T}} - 1\beta^{\mathrm{d}}_{i,l,d}(k)}, \tag{39}$$

where $\widetilde{k}$ is the bin index; $\widetilde{T}$ is the bin size (in this case, 30 min); D is the set of days considered, which, depending on the chosen strategy, may include all available previous days or just a subset of them (e.g. only the days corresponding to the same weekday, as per case a defined above); $b^{\mathrm{d}}_{i,l,d}(k)$ is the number of passengers boarding a vehicle of line l at station i during $(k-1,k]$ in day d; $t^{\mathrm{d}}_{i,l,d}(k)$ is the time elapsed between when the previous vehicle of line l stopped at station i and current time k; and $\beta^{\mathrm{d}}_{i,l,d}(k)$ is binary variable indicating if a run of line l departing from station i during $(k-1,k]$ on day d provides passenger load information.

Correspondingly, $\widetilde{\gamma}^{\mathrm{s}}_{i,l}(k)$ is calculated as

$$\widetilde{\gamma}^{\mathrm{s}}_{i,l}\left(\widetilde{k}\right) = \frac{\sum_{d\in D}\sum_{\widetilde{k=kT}}^{\widetilde{(k+1)}\widetilde{T}} - 1\overline{\gamma}^{d}_{i,l,d}(k)}{\sum_{d\in D}\sum_{\widetilde{k=kT}}^{\widetilde{(k+1)}\widetilde{T}} - 1\beta^{\mathrm{d}}_{i,l,d}(k)}, \tag{40}$$

where $\overline{\gamma}^{d}_{i,l,d}(k)$ is the alighting rate of vehicles of line l at station i during $(k-1,k]$ in day d.

4.4. Passenger comfort levels

The Level of Service (LOS) framework of Chandakas (2009) is used to measure in-vehicle comfort on-board a tramway vehicle, which assigns a comfort level on the basis of the on-board density of standing passengers (D) and of the availability of occupied seats (R), as shown in Table 1.

The in-vehicle comfort level is matched to specific locations (or stops) at specific times of the day. Fig. 7 illustrates the average LOS of the eastbound (left) and westbound (right) Line 1 per time step of the historical data. LOS 4, where standing takes place in tolerable conditions, is observed in the densest section of the central trunk of the line throughout the day, but extends to the start and end of the line during the morning, midday and afternoon peak hours. LOS 5 conditions are also observed sporadically during peak times.

Table 1
In-vehicle comfort LOS (Chandakas, 2009).

LOS	Conditions	Description
6	$D > 4$	Density greater than 4 standees per m^2, exceptional situation
5	$1.8 < D \leq 4$	No tolerance from standees, reliability issues
4	$0 < D \leq 1,8$	Standing conditions tolerated
3	$0.5 < R \leq 1$	Possibility to travel seated, limited choice available
2	$0.25 < R \leq 0.5$	Possibility to choose a seat, constrained to availability
1	$R \leq 0,25$	Free choice of a seat

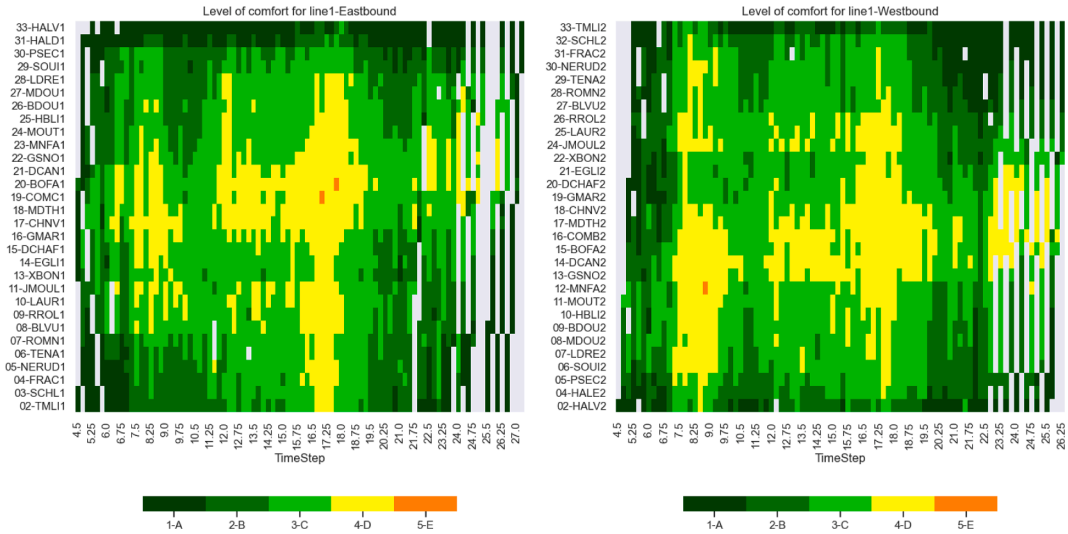

Fig. 7. In-vehicle comfort LOS for eastbound (left) and westbound (right) Line 1 of the Nantes tramway (the direction of the run is from bottom to top).

5. Evaluation and validation

This section documents the validation and evaluation process and results of the developed estimation method for real-time on-board passenger loads on the basis of AVL and APC data, based on its application on the Nantes tramway network. The evaluation metrics and methodology employed are presented first, and then the results obtained are reported and discussed.

5.1. Evaluation metrics and methodology

The proposed method is particularly aimed at estimating the passenger load for the runs that do not feature an APC system. However, as information is not available for such runs, an alternative validation approach is developed, which employs only data from runs equipped with APC. Specifically, when the estimation algorithm processes an APC-equipped run, estimation is first performed assuming that no APC measurement is available, denoting such estimate $\check{\omega}$ (for a generic variable ω) and storing the respective estimated values; then, the estimates are re-calculated considering the available APC measurements, and these latter values are then used for the subsequent estimation steps. In this way, the estimation method can be evaluated in a fair manner through comparison of the estimates obtained without considering APC measurements ($\check{\omega}$) with their corresponding actual measurements ($\overline{\omega}$).

In order to evaluate the estimation performance and assess its effectiveness, a series of experiments are conducted and different metrics are calculated. First, the main components of the proposed methodology are evaluated, namely the boarding and alighting estimators described in Sections 3.3 and 3.4. The mean absolute error (MAE) metric is employed in this respect, defined as:

$$\mathrm{MAE} = \frac{\sum_{k=0}^{K}\sum_{i\in I}\sum_{l\in L}\left|\check{\omega}^{\,s}_{i,l}(k) - \overline{\omega}^{\,s}_{i,l}(k)\right|}{\sum_{k=0}^{K}\sum_{i\in I}\sum_{l\in L}\beta^{s}_{i,l}(k)}, \tag{41}$$

where K is the estimation horizon, i.e. an operational day; and $\omega^{\,s}_{i,l}$ represents the variables considered for evaluation, which, in this case, are $b^{s}_{i,l}(k)$, $\gamma^{s}_{i,l}(k)$, and (indirectly) $a^{s}_{i,l}(k)$. In addition, the weighted mean absolute percentage error (WMAPE) is employed as a relative metric, defined as:

$$\mathrm{WMAPE} = \frac{\sum_{k=0}^{K}\sum_{i\in I}\sum_{l\in L}\left|\check{\omega}^{\,s}_{i,l}(k) - \overline{\omega}^{\,s}_{i,l}(k)\right|}{\sum_{k=0}^{K}\sum_{i\in I}\sum_{l\in L}\left|\overline{\omega}^{\,s}_{i,l}(k)\right|}, \tag{42}$$

This metric overcomes the infinite error issue of other relative metrics (such as the mean absolute percentage error, MAPE); however, it is still expected to suffer from the limitation of over-penalising cases when real values are low.

The focus then shifts to the estimated passenger comfort, which is calculated according to the method described in Section 4.4, assuming such information is available every time a vehicle stops at a station. In this case, the sequence of estimated comfort levels (again, assuming no APC measurement was available) is compared with the sequence of comfort level calculated from actual measured data. In this comparison, besides investigating aggregated measures such as the mean absolute comfort error, more disaggregated information is further considered, such as the distribution of errors for different error classes.

5.2. Results and discussion

The performance of the individual components of the estimation method is first investigated, i.e. the boarding and the alighting rate estimators. The estimation results for the estimator of boarding passengers are presented in Tables 2 and 3. As can be observed, the estimation MAEs and WMAPEs for all days investigated are overall low; all MAEs are lower than 10 passengers and most WMAPEs remain below 1 (bearing in mind that they are also affected by the low actual values). Moreover, it can be seen that the performance of the estimator improves by incorporating historical data, with a maximum improvement of the order of 25 % on Friday 20 December 2019 in both the MAE and the WMAPE values. In addition, it can be observed that the case where the estimator is fed with historical data from all days outperforms the case that employs historical data only from the same weekdays. This suggests that the assumption that different days may have different peak hour patterns (which is the reason why the use of historical data only from the same weekdays was proposed in the first place) is uninfluential and that its effects are most probably outweighed by the fact that greater data availability allows obtaining better results. A notable exception is Thursday 12 December 2019, which is also characterised by considerably higher MAE and WMAPE values than the other days; a deeper investigation into that day, though, has revealed that a service disruption lasting about 40 min that happened during the afternoon has caused the estimator to produce some unreasonably high errors.

The performance of the alighting passengers' estimator is demonstrated in Tables 4 and 5, where a similar pattern as the one observed for the boarding passengers' estimator can be observed. On the other hand, the MAE and WMAPE values calculated for the alighting rates without using historical data already look reasonably low. Still, even in this case the use of historical data improves the estimation performance even further and, again, best results are obtained when all days are used to assimilate the historical data.

The results in terms of passenger comfort (absolute) errors are investigated next, calculated as described in Section 4.4. The results are shown in Table 6, where it can be observed that the resulting comfort levels obtained via the estimation scheme are accurate, with an absolute error lower than one comfort level for all examined scenarios, except one (the service disrupted impacted Thursday 12 December 2019, without using historical data). From these results, the value of employing historical data in addition to real-time measurements is highlighted again.

Considering the passenger comfort estimation results in more detail, Fig. 8 shows how the comfort errors (calculated as estimated level minus measured comfort level) are distributed. Firstly, it is interesting to observe that the number of data points with a comfort error equal to zero have by far the highest count, reaching almost 50 % of all observations for the case when historical data for all days are used. Secondly, it can be seen that the occurrence of error points decreases as the (absolute) error increases, i.e., the second most frequent errors are ± 1, while higher errors (e.g. 4 and 5) are either extremely rare or even completely absent. Thirdly, the use of historical data not only increases the average performance, as discussed earlier, but also makes the distribution of errors narrower, meaning fewer higher error occurrences.

Moreover, it is interesting to mention that, from a practical viewpoint, a negative error (i.e. when a more comfortable condition than the actual one is estimated) is more critical than a positive error (i.e. when a less comfortable condition than the actual one is estimated), since the former would tell end-users that a run is less crowded than what it is in reality, which may, in turn, lead to poor decisions (e.g. even more crowding in an already crowded run). In this respect, it is interesting to observe that by using historical data, the samples with negative error diminish more than the ones with positive error.

Furthermore, Figs. 9 and 10 present an example of disaggregated information on the passenger comfort estimation for a selected day. In particular, Fig. 9 shows contour plots for the measured comfort, the estimated comfort, and the comfort error, calculated for all the runs equipped with APC for Wednesday 18 December 2019, and where estimates are produced without utilising historical data.

Table 2
MAE for the estimated number of boarding passengers, employing different historical data settings.

Day	No historical data ($\varphi = 0$)	Historical data only from same weekdays	Historical data from all days
Wed 11 Dec	6.411	5.768	5.320
Thu 12 Dec	9.100	9.484	9.327
Fri 13 Dec	6.282	6.422	5.696
Mon 16 Dec	7.050	6.184	6.147
Wed 18 Dec	5.458	5.214	4.351
Thu 19 Dec	7.071	7.094	6.497
Fri 20 Dec	7.944	6.696	5.914

Table 3
WMAPE for the estimated number of boarding passengers, employing different historical data settings.

Day	No historical data ($\varphi = 0$)	Historical data only from same weekdays	Historical data from all days
Wed 11 Dec	0.881	0.803	0.733
Thu 12 Dec	1.209	1.217	1.271
Fri 13 Dec	0.726	0.689	0.621
Mon 16 Dec	0.959	0.950	0.935
Wed 18 Dec	0.773	0.760	0.658
Thu 19 Dec	0.854	0.735	0.801
Fri 20 Dec	1.033	0.835	0.756

Table 4
MAE for the estimated alighting rate and number of alighting passengers, using different historical data settings.

Day	Alighting rate			Alighting passengers		
	No historical data ($\varphi = 0$)	Historical data only from same weekdays	Historical data from all days	No historical data ($\varphi = 0$)	Historical data only from same weekdays	Historical data from all days
Wed 11 Dec	0.106	0.099	0.092	7.819	7.810	7.816
Thu 12 Dec	0.091	0.089	0.075	7.610	7.613	7.642
Fri 13 Dec	0.092	0.083	0.078	9.155	9.145	9.149
Mon 16 Dec	0.099	0.085	0.077	8.022	8.016	8.041
Wed 18 Dec	0.092	0.086	0.078	7.613	7.632	7.618
Thu 19 Dec	0.079	0.073	0.071	8.594	8.611	8.612
Fri 20 Dec	0.090	0.090	0.076	8.689	8.682	8.678

Table 5
WMAPE for the estimated alighting rate and number of alighting passengers, using different historical data settings.

Day	Alighting rate			Alighting passengers		
	No historical data ($\varphi = 0$)	Historical data only from same weekdays	Historical data from all days	No historical data ($\varphi = 0$)	Historical data only from same weekdays	Historical data from all days
Wed 11 Dec	0.633	0.573	0.555	0.940	0.945	0.946
Thu 12 Dec	0.552	0.475	0.496	0.945	0.956	0.946
Fri 13 Dec	0.550	0.493	0.470	0.939	0.934	0.934
Mon 16 Dec	0.608	0.536	0.569	0.944	0.956	0.956
Wed 18 Dec	0.608	0.555	0.525	0.944	0.962	0.951
Thu 19 Dec	0.558	0.517	0.514	0.912	0.921	0.917
Fri 20 Dec	0.570	0.536	0.504	0.937	0.937	0.933

Table 6
Absolute comfort level error employing different settings for historical data.

Day	No historical data ($\varphi = 0$)	Historical data only from same weekdays	Historical data for all days
Wed 11 Dec	0.859	0.744	0.734
Thu 12 Dec	1.031	0.715	0.712
Fri 13 Dec	0.694	0.630	0.588
Mon 16 Dec	0.837	0.651	0.692
Wed 18 Dec	0.864	0.710	0.646
Thu 19 Dec	0.728	0.609	0.564
Fri 20 Dec	0.821	0.553	0.504

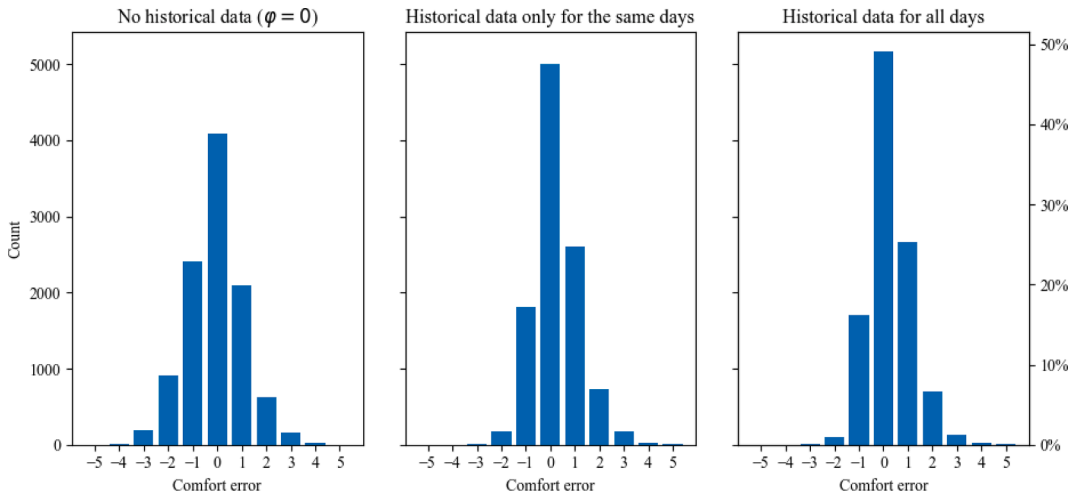

Fig. 8. Distribution of the comfort errors for all days, considering a total of 10,523 data points.

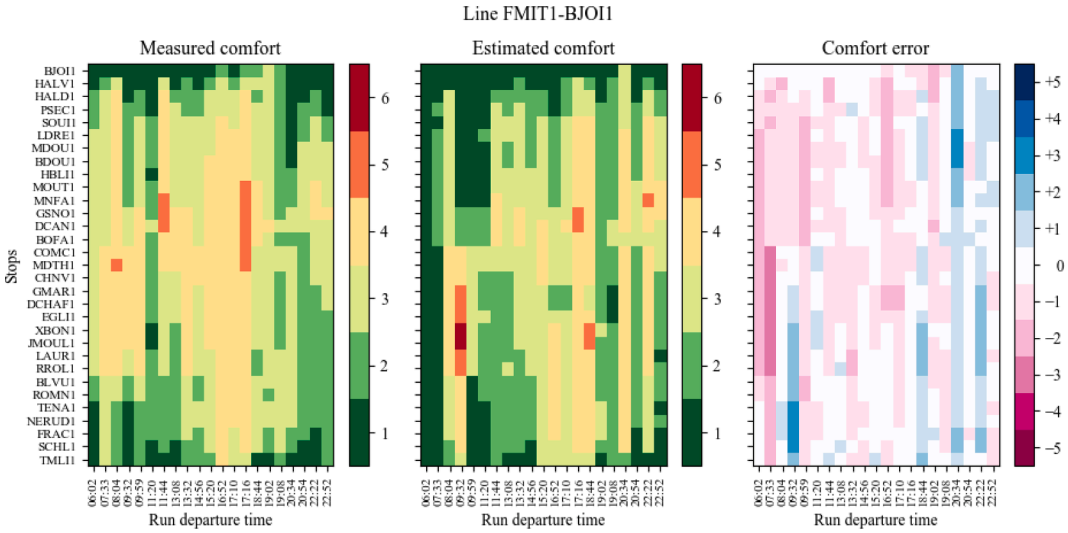

Fig. 9. Measured comfort level (left); its estimates using the proposed method (centre); and the estimation error calculated in terms of comfort error (right), for Wednesday 18 December 2019, without utilising historical data.

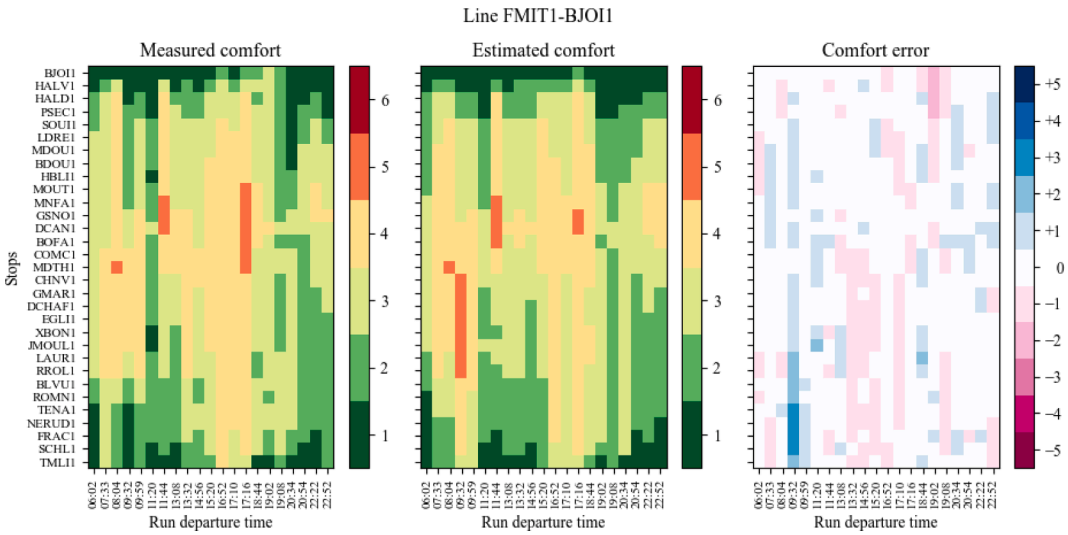

Fig. 10. Measured comfort level (left); its estimates using the proposed method (centre); and the estimation error calculated in terms of comfort error (right), for Wednesday 18 December 2019, utilising historical data for all days.

Correspondingly, Fig. 10 shows the same contour plots, but in this case estimates are produced utilising historical data for all days. Again, the good performance of the proposed estimation scheme is demonstrated, as generally very low errors are produced, while it can be clearly seen that, in the case where historical data for all days are utilised, large areas in the contour plot are white (denoting zero comfort error) and swathes of purple and pink coloured cells (denoting negative comfort errors) are eliminated. It can also be observed how, at the disaggregated level, the negative errors reduce both in total amount and in magnitude when utilising historical data, which is therefore a preferable choice.

The findings obtained are further supported by the graphs in Fig. 11, which show the distribution of the comfort errors for all days, grouped by the measured comfort levels. As can be seen, the impact of the use of historical data in terms of the estimation accuracy becomes most prominent at the higher levels of on-board crowding (namely levels 3 and 4), where most of the severe negative discrepancies (of two or three levels) are either completely eliminated or, at least, attenuated. Indeed, just like in the example day of Figs. 9 and 10, the pink and purple areas are severely diminished as more historical data are employed. This suggests that the proposed estimation scheme is not only able to reliably estimate the actual on-board comfort level with only few errors, but is also capable of ensuring that any remaining errors are low and not critical (i.e. no instances of a crowded vehicle estimated as empty).

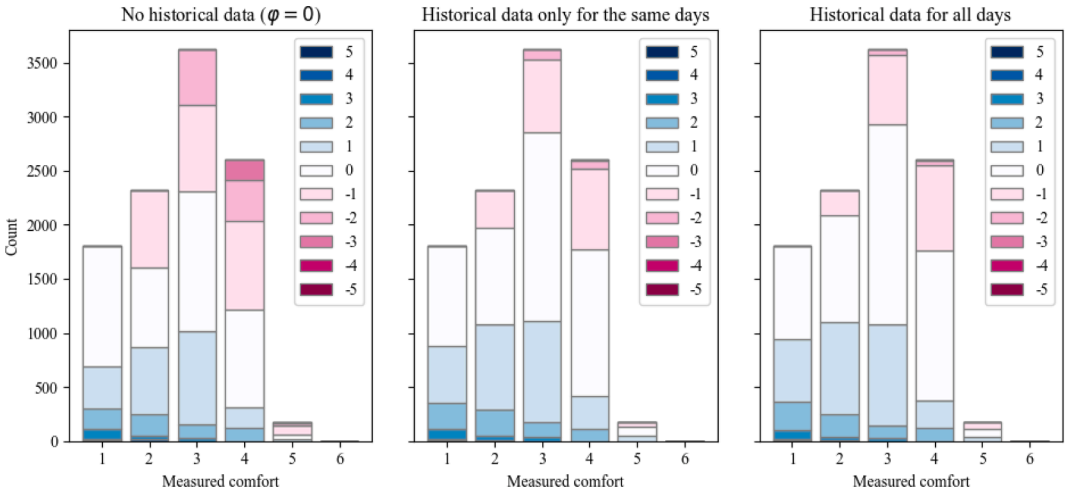

Fig. 11. Distribution of the comfort errors for all days, grouped by the measured comfort levels.

Furthermore, it is investigated how the comfort error metric changes with the amount of historical data available to the estimation methodology, and to this end, experiments are performed by utilising an increasing number of days from the set of available historical data, from 5 to 34 days (i.e. all available days), and evaluating the performance for the same days as tested in the previous experiments. The results are presented in Fig. 12, where it can be observed that as more historical data become available, the estimation performance improves. It is particularly interesting to highlight that, after a certain amount of historical data are available, adding further data improves the estimation performance only marginally; in the experiments carried out in this study, for example, this happens when more than 15 days of data are available.

Finally, it is investigated how the performance of the proposed estimator varies with different amounts of APC data. These experiments are conducted by randomly selecting a percentage of the available APC equipped runs and performing estimation so that the proposed estimator employs only APC data from the selected runs. In particular, it is tested how the estimator performs using 25 %, 50 % and 75 % of the available APC data, which roughly corresponds to 2.5 %, 5 % and 7.5 % of the total number of daily runs, respectively (as per Section 4.2). The same percentages are used to sample the data treated as historical, as well as the data treated as real-time. In order to reduce noise due to the stochastic nature of the selection, five different random seeds are employed, and the arithmetic averages of the estimation metrics obtained for five replications for each of the tested percentages are reported. The results are presented in Fig. 13, where the case where all the available APC data are used (i.e. 100 %) is also shown. As can be observed, and as can be expected, in virtually all tested days the estimation performance improves by increasing the amount of APC data available, with improvements being more significant at lower levels of APC data availability (i.e. from 25 % to 50 %).

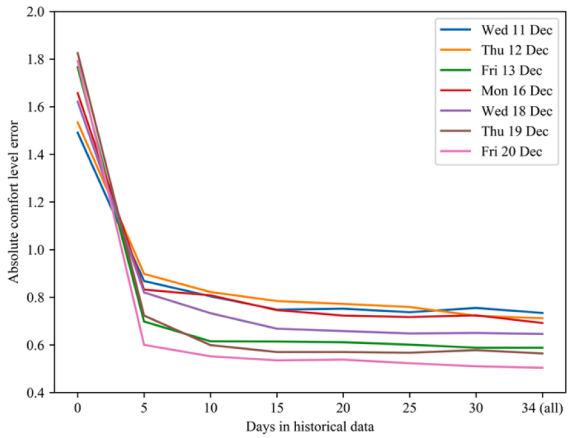

Fig. 12. Absolute comfort level error obtained via utilising different amounts of historical data.

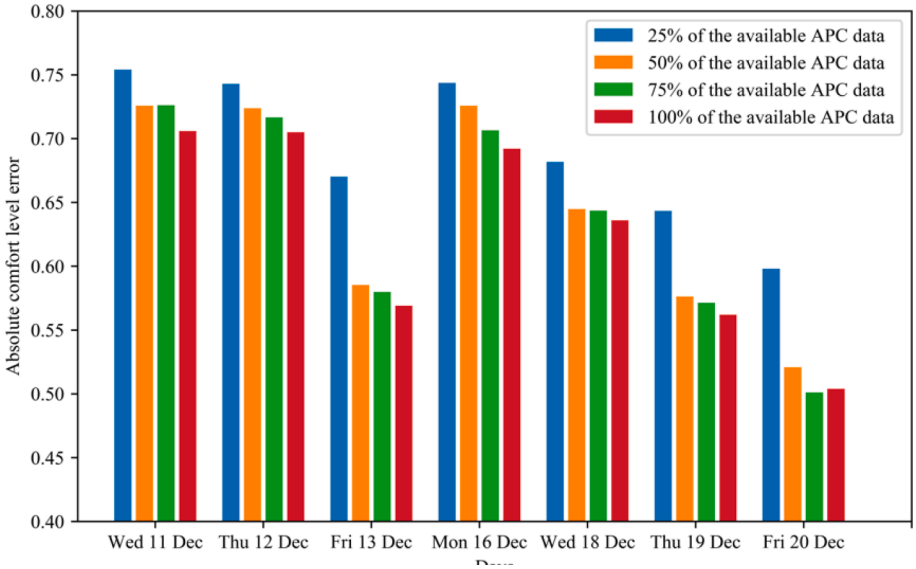

Fig. 13. Absolute comfort level error obtained by utilising different amounts of APC data.

6. Conclusions

The provision of accurate and reliable passenger comfort information on-board public transport vehicles has long been identified as an important constituent component of good customer service. In light of the COVID-19 pandemic and its associated impacts on public transport patronage, however, the provision of such information has become an absolute necessity for the "survival" of public transport services going forward. The present study, therefore, has proposed a new KF-based estimation scheme for on-board comfort levels, which employs historical and current non-exhaustive APC data, as well as AVL measurements. The method has been successfully validated through application to the tramway network of the French city of Nantes, with the results demonstrating that overall low estimation errors are delivered, which are also non-critical.

But while the study has made an initial contribution to the topic on on-board passenger comfort estimation, research in this direction continues. For instance, a limitation of the present study is that, as identified in Section 3.5, the system is not always fully observable. This means that there are some caveats in the estimation, such as in early-morning runs (due to the non-availability of previous measurements (apart from historical data)) and conditions with high passenger flow volatility (e.g. service disruptions). Indeed, it can be observed from the experiments of this study that the vast majority of the (remaining) errors originate in such situations, and it is the objective of future work to address this aspect, by including additional data sources, such as WiFi and mobile data. The development of these device-based capturing techniques could impact the necessity and the extent to which APC systems are deployed. In addition, it would be interesting to investigate how the frequency and distribution of APC-equipped vehicles affect estimation performance, which could be properly evaluated if a larger amount of APC data were available. To this extent, further analysis could explore how the estimation performance of the proposed scheme would compare with that of other methods, such as that of Jenelius (2019), and how it could be potentially improved by combining and integrating these methods. An alternative approach could involve using data produced by a simulator, which would allow randomly determining whether any (simulated) run is APC-equipped or not (naturally with limitations in terms of the realism of the demand and arrival patterns).

Moreover, a limitation of the present study is that the proposed estimation method has so far not been applied in a real-time context, but only in an "offline" one. The reason for this is not inherent to the estimation method itself, but rather the fact that the two systems used in the validation case study (APC and AVL) operate independently from each other and their data are therefore not synchronised. A considerable amount of data pre-processing has therefore been necessary here in order to assimilate both the historical and the current datasets. In order to be able to perform real-time on-board comfort estimation, synchronisation of the APC and AVL data feeds using a vehicle label and a timestamp would be required so as to improve the run-matching process. In practice this is usually enabled by an architecture with continuous or semi-continuous communication between the systems. It would be, hence, useful to explore how this could be achieved in the current case study, or to apply the method in a real-time context on a different case study. This would additionally enable the monitoring of passenger behaviour and perceptions in response to the information provided by the estimation scheme.

Finally, further research will concentrate on extending the capabilities and accuracy of the proposed method. For example, it would be interesting to investigate the performance of the estimation scheme in other sites, featuring more complex network topologies and a greater variety of modes, and to also extend the remit of the method from estimation (at the current station and time) to prediction (at a later station and/or future time point). Moreover, additional scenarios covering the phenomena of overcrowding and left-behind passengers (which were not present in the dataset employed here), may be further studied; these will likely require some special

treatments and potentially also some modifications to the proposed method. Such extensions could provide a much different insight into how on-board comfort information would influence a priori and real-time passenger planning decisions in multi-modal trips and would consequently allow modelling the impacts at the route/mode choice and assignment levels.

CRediT authorship contribution statement

Claudio Roncoli: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Ektoras Chandakas:** Conceptualization, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Ioannis Kaparias:** Conceptualization, Investigation, Resources, Visualization, Writing – original draft, Writing – review & editing, Project administration.

Declaration of Competing Interest

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

Acknowledgements

The authors would like to thank Semitan for supplying the data used in this study, and in particular the ‘‘Direction de la Performance et de l’Innovation’’ for their assistance in analysing the Autumn 2019 Opthor and Ineo datasets. The author Claudio Roncoli also acknowledges the support of the Academy of Finland project ALCOSTO (349327).

References

Achar, A., Bharathi, D., Kumar, B.A., Vanajakshi, L., 2020. Bus arrival time prediction: a spatial Kalman filter approach. IEEE Trans. Intell. Transp. Syst. 21, 1298–1307.
Anderson, B.D.O., Moore, J.B., 1979. Optimal Filtering. Prentice-Hall.
Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., 2010. Kalman Filter applications for traffic management. In: Kordic, V. (Ed.), Kalman Filter. IntechOpen.
Antsaklis, P., Michel, A.N., 2006. Linear Systems. Birkhäuser.
Bai, Y., Sun, Z., Zeng, B., Deng, J., Li, C., 2017. A multi-pattern deep fusion model for short-term bus passenger flow forecasting. Appl. Soft Comput. 58, 669–680.
Batarce M., Muñoz J.C., Ortúzar J. de D., Raveau S., Mojica C., Ríos A., 2015. Valuing crowding in public transport systems using mixed SP/RP data: the case of Santiago. Transport. Res. Record 2535, 73-78.
Bekiaris-Liberis, N., Roncoli, C., Papageorgiou, M., 2016. Highway traffic state estimation with mixed connected and conventional vehicles. IEEE Trans. Intell. Transp. Syst. 17, 3484–3497.
Bekiaris-Liberis, N., Roncoli, C., Papageorgiou, M., 2016. Highway traffic state estimation per lane in the presence of connected vehicles. Transp. Res. B 106, 1–28.
Chandakas E., 2009. La capacité des transports ferroviaires d’Ile-de-France face à la hausse du trafic à long-terme. Master Thesis, Ecole Nationale des Ponts et Chaussées.
Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10.
Cox, T., Houdmont, J., Griffiths, A., 2006. Rail passenger crowding, stress, health and safety in Britain. Transp. Res. A 40, 244–258.
De Palma, A., Kilani, M., Proost, S., 2015. Discomfort in mass transit and its implication for scheduling and pricing. Transp. Res. B 71, 1–18.
De Solere, R. (Ed), 2012. La mobilité urbaine en France: Enseignements des années 2000-2010. Editions du CERTU, Collection Références.
Ding, C., Duan, J., Zhang, Y., Wu, X., Yu, G., 2018. Using an ARIMA-GARCH modeling approach to improve subway short-term ridership forecasting accounting for dynamic volatility. IEEE Trans. Intell. Transp. Syst. 19, 1054–1064.
European Committee for Standardization (CEN) (2002). *EN 13816:2002 Transportation – Logistics and services – Public passenger transport – Service quality definition, targeting and measurement*, CEN/TC 320 - Transport - Logistics and services. Available.
Gkiotsalitis, K., Cats, O., 2021. Public transport planning adaption under the COVID-19 pandemic crisis: literature review of research needs and directions. Transp. Rev. 41, 374–392.
Gong, M., Fei, X., Wang, Z.H., Qiu, Y.J., 2014. Sequential framework for short-term passenger flow prediction at bus stop. Transp. Res. Rec. 2417, 58–66.
Guo, J., Huang, W., Williams, B.M., 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. Transp. Res. C 43, 50–64.
Gur, Y.J., Ben-Shabat, E., 1997. Estimating bus boarding matrix using boarding counts in individual vehicles. Transp. Res. Rec. 1607, 81–86.
Haywood, L., Koning, M., 2013. Estimating crowding costs in public transport. DIW Berlin Discussion Paper No. 1293.
Haywood, L., Koning, M., Monchambert, G., 2017. Crowding in public transport: Who cares and why? Transp. Res. A 100, 215–227.
He, H., Yan, M., Sun, C., Peng, J., Li, M., Jia, H., 2018. Predictive air-conditioner control for electric buses with passenger amount variation forecast. Appl. Energy 227, 249–261.
Heydenrijk-Ottens, L., Degeler, V., Luo, D., van Oort, N., van Lint, H., 2018. Supervised learning: Predicting passenger load in public transport. Conference on Advanced Systems in Public Transport (CASPT) 2018.
Hu, R., Chiu, Y.-C., Hsieh, C.-W., 2020. Crowding prediction on mass rapid transit systems using a weighted bidirectional recurrent neural network. IET Intel. Transport Syst. 14, 196–203.
Huang, D., Yu, J., Shen, S., Li, Z., Zhao, L., Gong, C., 2020. A method for bus OD matrix estimation using multisource data. J. Adv. Transp. 2020, 5740521.
Jenelius, E., 2020. Data-driven metro train crowding prediction based on real-time load data. IEEE Trans. Intell. Transp. Syst. 21, 2254–2265.
Jenelius, E. (2019). ‘‘Data-driven bus crowding prediction based on real-time passenger counts and vehicle locations’’. 6th International Conference on Models and Technologies for Intelligent Transport Systems (MT-ITS 2019).
Ji, Y., Mishalani, R.G., McCord, M.R., 2015. Transit passenger origin–destination flow estimation: Efficiently combining onboard survey and large automatic passenger count datasets. Transp. Res. C 58, 178–192.
Jia, F., Li, H., Jiang, X., Xu, X., 2019. Deep learning-based hybrid model for short-term subway passenger flow prediction using automatic fare collection data. IET Intel. Transport Syst. 13, 1708–1716.
Jung, Y.J., Casello, J.M., 2020. Assessment of the transit ridership prediction errors using AVL/APC data. Transportation 47, 2731–2755.
Kalman, R.E., Bucy, R.S., 1961. New results in linear filtering and prediction theory. Transactions of the ASME Series D 83, 95–108.
Khomchuk, P., Tuladhar, S.R., Sivananthan, S., 2018. Predicting passenger loading level on a train car: A Bayesian approach. arXiv eprints, p. 1808.06962.

Koutsopoulos, H.N., Noursalehi, P., Zhu, Y., Wilson, N.H.M., 2017. Automated data in transit: Recent developments and applications. 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS 2017), Naples, Italy.
Koutsopoulos, H.N., Ma, Z., Noursalehi, P., Zhu, Y., 2019. Chapter 10 - Transit data analytics for planning, monitoring, control, and information. In: Actoniou, C., Dimitriou, L., Pereira, F. (Eds.), Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modeling. Elsevier, pp. 229–261.
Kroes, E., Kouwenhoven, M., Debrincat, L., Pauget, N., 2013. "On the value of crowding in public transport for Île-de-France. International Transport Forum Discussion Papers, no. 2013/18, Organisation for Economic Co-operation and Development (OECD).
Leurent, F., Liu, K., 2009. On seat congestion, passenger comfort and route choice in urban transit: a network equilibrium assignment model with application to Paris, 88th Annual Meeting of the Transportation Research Board, Washington DC, USA.
Leurent, F., Chandakas, E., Poulhes, A., 2013. A traffic assignment model for passenger transit on a capacitated network: Bi-layer framework, line sub-models and large-scale application. Transp. Res. C 47, 3–27.
Li, D., Lin, Y., Zhao, X., Song, H., Zou, N., 2011. Estimating a transit passenger trip Origin-Destination matrix using Automatic Fare Collection system. In: Xu, J., Yu, G., Zhou, S., Unland, R. (eds), Database Systems for Advanced Applications (DASFAA 2011), Lecture Notes in Computer Science, vol. 6637. Springer.
Li, Z., Hensher, D.A., 2013. Crowding in public transport: a review of objective and subjective measures. J. Public Transp. 16, 107–134.
Lin, C.-T., 1974. Structural controllability. IEEE Trans. Autom. Control 19, 201–208.
Liu, Y-Y., Slotine, J-J. , Barabasi, A-L., 2013. Observability of complex systems. Proceedings of the National Academy of Sciences, vol. 110, pp. 2460–2465.
Liu, X., van Hentenryck, P., Zhao, X., 2021. "Optimization models for estimating transit network Origin-Destination flows with big transit data. J. Big Data Anal. Transport. 3, 247–262.
Mikkelsen, L., Buchakchiev, R., Madsen, T., Schwefel, H.P., 2016. Public transport occupancy estimation using WLAN probing. 8th International Workshop on Resilient Networks Design and Modeling (RNDM), 302-308.
Mohammadi, A., Amador-Jimenez, L., Nasiri, F., 2020. A multi-criteria assessment of the passengers' level of comfort in urban railway rolling stock. Sustain. Cities Soc. 53, 101892.
Noursalehi, P., 2017. Decision Support Platform for Urban Rail Systems: Real-time Crowding Prediction and Information Generation, PhD thesis, Northeastern University, Boston, MA, USA.
Oransirikul, T., Nishide, R., Piumarta, I., Takada, H., 2014. Measuring bus passenger load by monitoring Wi-Fi transmissions from mobile devices. Procedia Technol. 18, 120–125.
Pasini, K., Khouadjia, M., Ganansia, F., Oukhellou, L., 2019. Forecasting passenger load in a transit network using data driven models". 12th World Congress on Railway Research (WCRR 2019).
Pasini, K., Khouadjia, M., Samé, A., Ganansia, F., Oukhellou, L., 2020. LSTM encoder-predictor for short-term train load forecasting. In: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (eds), Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2019, Lecture Notes in Computer Science, vol. 11908, pp 535-551. Springer.
Przybylowski, A., Stelmak, S., Suchanek, M., 2021. Mobility behaviour in view of the impact of the COVID-19 pandemic – Public transport users in Gdansk case study. Sustainability 13.
Reissig, G., Hartung, C., Svaricek, F., 2014. Strong structural controllability and observability of linear time-varying systems. IEEE Trans. Autom. Control 59, 3087–3092.
Roncoli, C., Bekiaris-Liberis, N., Papageorgiou, M., 2016. Use of speed measurements for highway traffic state estimation: case studies on NGSIM data and Highway A20, Netherlands. Transp. Res. Rec. 2559, 90–100.
Simon, D., 2010. Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. IET Control Theory Appl. 4, 1303–1318.
Simon, D., 2006. Optimal State Estimation: Kalman, H Infinity and Nonlinear Approaches. John Wiley & Sons.
Sun, W., Schmöcker, J.-D., Fukuda, K., 2021. Estimating the route-level passenger demand profile from bus dwell times. Transp. Res. C 130, 103273.
Szeto, M., Gazis, D., 1972. Application of Kalman Filtering to the surveillance and control of traffic systems. Transp. Sci. 6, 419–439.
Tao, Z., Tang, J., 2019. Real-time estimation of urban rail transit passenger flow status based on multi-source data. J. Phys. Conf. Ser. 1187, 052070.
Tirachini, A., Heshner, D.A., Rose, J.M. (2012). "Multimodal pricing and optimal design of public transport services: The interplay between traffic congestion and bus crowding". 12th International Conference on Advanced Systems of Public Transport, Santiago, Chile.
Tirachini, A., Hurtubia, R., Dekker, T., Daziano, R., 2017. Estimation of crowding discomfort in public transport: results from Santiago in Chile. Transp. Res. A 103, 311–326.
Transport Focus, 2021. "Will there be space on board?". Available: https://www.transportfocus.org.uk/publication/will-there-be-space-onboard.
Trozzi, V., Gentile, G., Bell, M.G.H., Kaparias, I., 2013. Dynamic user equilibrium in public transport networks with passenger congestion and hyperpaths. Transp. Res. B 57, 266–285.
Tsai, T.-H., Lee, C.-K., Wei, C.-H., 2009. Neural network based temporal feature models for short-term railway passenger demand forecasting. Expert Syst. Appl. 36, 3728–3736.
Turner, S., Corbett, E., O'Hara, R., White, J., 2005. Health and safety effects of rail crowding: hazard identification. RSSB R&D Programme. T307 Report,.
US Transportation Research Board, 2013. Transit Capacity and Quality of Service Manual, Third Edition. The National Academies Press.
van Oort, N., Brands, T., de Romph, E., 2015. Short-term prediction of ridership on public transport with smart card data. Transp. Res. Rec. 2535, 105–111.
Vlahogianni, E., Golias, J.C., Karlaftis, M.G., 2004. Short-term traffic forecasting: Overview of objectives and methods. Transp. Rev. 24, 533–557.
Vlahogianni, E., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: where we are and where we're going. Transp. Res. C 43, 3–19.
Wang, W., Attanucci, J., Wilson, N., 2011. Bus passenger Origin-Destination estimation and related analyses using automated data collection systems. J. Public Transp. 14, 131–150.
Wang, P., Chen, X., Chen, J., Hua, M., Pu, Z., 2021. A two-stage method for bus passenger load prediction using automatic passenger counting data. IET Intel. Transport Syst. 15, 248–260.
Wang, Y., Papageorgiou, M., 2005. Real–time freeway traffic state estimation based on extended Kalman filter: A general approach. Transp. Res. B 39, 141–167.
Zhang, Y., Jenelius, E., Kottenhoff, K., 2017. Impact of real-time crowding information: a Stockholm metro pilot study. Public Transport 9, 483–499.
Zhang, J., Shen, D., Tu, L., Zhang, F., Xu, C., Wang, Y., Tian, C., 2017. A real-time passenger flow estimation and prediction method for urban bus transit systems. IEEE Trans. Intell. Transport. Syst. 18, 3168–3178.