Author/s:
Yin, T;Nassir, N;Leong, J;Tanin, E;Sarvi, M

Title:
Transferable supervised learning model for public transport service load estimation

Date:
2023-07-28

Citation:
Yin, T., Nassir, N., Leong, J., Tanin, E. & Sarvi, M. (2023). Transferable supervised learning model for public transport service load estimation. Transportation, 52 (1), pp.29-54. https://doi.org/10.1007/s11116-023-10411-2.

Persistent Link:
https://hdl.handle.net/11343/336641

Transportation
https://doi.org/10.1007/s11116-023-10411-2

Transferable supervised learning model for public transport service load estimation

Tianwei Yin[1] · Neema Nassir[1] · Joseph Leong[1] · Egemen Tanin[2] · Majid Sarvi[1]

Accepted: 17 July 2023

Abstract
Detailed knowledge of service utilisation and passenger load profiles is the basis for the design, operation, and adjustment of a public transport service. The advancement in sensing technologies enable transit operators to monitor the variabilities in passenger flows continuously and consistently. There is a growing body of literature on using supervised learning models with direct passenger counts from historical observations. However, the incomplete, inaccurate, and biased data from automatic sensors pose challenges in this process. This paper proposes novel supervised learning models to estimate the onboard load profile of public transport services based on two main data sources: (1) limited data collected on a subset of service vehicles by automatic passenger counting (APC) systems, and (2) fare data collected by automatic fare collection (AFC) systems. The specific consideration is given to the fact that the developed models can be transferred across different routes. This is motivated by the commonly "limited coverage" of automated passenger counter devices on service vehicles. We introduce an array of new models, including a superior segment-based model, which demonstrates remarkable improvement in model transferability and accuracy. The proposed methodology utilises separate methods in different segments of a transit line. The proposed models were applied to three tram lines in Melbourne, Australia, where various types of shortcomings exist in the automated data. The test results demonstrate that the proposed models can be transferred and applied to other transit route without relying on historical observations. This would enable transit operators to reduce the number of required devices and monitor service utilisation in a more cost-efficiently manner, particularly in public transport networks where AFC coverage is usually incomplete and negatively skewed. The information on service utilisation will not only help operators to accommodate the variability in passenger demand but also assist passengers in journey planning to avoid overcrowding on services.

Keywords Passenger loads estimation · Supervised learning · Transfer learning · Missing data

✉ Tianwei Yin
tiyin@student.unimelb.edu.au

[1] Department of Infrastructure Engineering, The University of Melbourne, Melbourne, Australia

[2] School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia

Published online: 28 July 2023

Introduction

Rapid urbanisation is placing increasing stress on transportation infrastructure. The estimation of passenger flows is the basis for the design, operation, and adjustment of a public transport network. For applications in strategic planning, a fundamental understanding of regular travel demand volumes at zone-aggregated levels is usually sufficient. This information can be accessed through the four-step travel demand forecasting model (McNally 2007). However, for operational applications, demand variability information is becoming increasingly relevant and essential.

The transit industry has recently started to utilise demand information to support data-driven decision making. For instance, transit operators can adjust service frequency to accommodate the variability in passenger demand, develop demand-responsive services to target areas with low passenger demand, provide customers with information on crowding levels, and assist them in their journey planning to avoid service overcrowding. The development of these strategies requires a deeper understanding of detailed passenger demand information, including day-to-day and within-day variability, as compared to strategic-level models. It is particularly important to understand demand responses to service disruptions, special events, restrictions (such as COVID-19) and operational interventions (through "before/after" analysis). Providing transit operators with demand patterns during these events can enhance their ability to respond to uncertainty and provide customers with more reliable and efficient services. However, conventional methods of collecting demand data through on-board surveys are labour-intensive and costly. As a result, sample sizes for operational applications are inevitably small.

Most transit operators recognise that data generated by Transit Intelligent Transportation Systems (ITS), including Automatic Passenger Counting (APC) and Automatic Fare Collection (AFC) systems, contain valuable information to support operational decision making. APC systems record detailed boarding and alighting demand by stop and time of day. However, due to budget limitations, APC systems are usually installed on a subset of fleet vehicles, making it impossible to directly monitor passenger loads for all individual trips. AFC systems were initially designed for financial and revenue purposes but are being recognised as a rich data source for understanding travel demand patterns. There is a growing literature that seeks to derive origin–destination flow matrices from AFC data (Trépanier et al. 2007; Pelletier et al. 2011; Gordon et al. 2013; Nassir et al. 2015, 2017, 2019; Munizaga et al. 2020) and analyse transit ridership patterns. This data is often more accessible than the direct passenger counts collected by APC systems. However, many cities lack access to high-quality AFC data due to missing pieces of information resulting from non-card users, free service zones, possible fare evasions etc. Therefore, a challenging problem that arises in this domain is the low coverage of high-quality data.

In this paper, our objective is to utilise data sources that are accessible to most transit operators and develop a data-driven model for reliable estimates of load profiles for every service trip, particularly those without direct passenger count data. We propose a set of supervised learning models using the ground-truth label derived from existing observations collected by APC systems. Considering the coverage limitations of APC systems, we address two problems: (1) same-route estimation, where supervised learning models are trained and tested using data from the same transit route, and (2) cross-route estimation, where a model trained with data from routes with direct measurement is applied to a new route without any historical observations. Both problems are examined using three tram lines in Melbourne, Australia, where various shortcomings exist in the passively collected

data. For instance, only 3 out of 24 routes are equipped with APC systems. In addition, concerning AFC data, fare evasion rates are relatively high, and a free tram zone is located in the Central Business District (CBD), where passengers are not required to tap on or off tram services.

The case study results demonstrate that supervised learning models generate more accurate estimates of passenger loads compared to direct estimation from AFC records. An important finding is that model accuracy can be enhanced by integrating supervised learning with direct correlation measures between boarding (or alighting) flows and on-board loads. Another significant finding, particularly in relation to cross-route estimation, is that a segment-based estimation method, which applies separate methods to different segments of a transit route, can yield superior transferability. This method is developed based on the hypothesis that the relationships between passenger flows and independent explanatory variables may not always consistent across different route segments. For upstream stops in the inbound direction, where most trips end up in the CBD, boarding flows are likely to be most affected by trip production features (such as residential population) and are expected to be consistent across different routes. However, passenger alighting flows may be primarily affected by the purposes of trips, which can significantly vary across different routes. On the other hand, for mid-route and downstream stops, boarding passengers do not have common destinations, so the boarding flows are not dependent on shared features across different routes. Hence, instead of utilising supervised learning model with uniform parameters along the routes, this method employs separate estimation methods to estimate boarding and alighting flows for different route segments/directions.

To the best of our knowledge, this paper is the first attempt to apply supervised learning models for transferable estimation of passenger flows, where training and test data come from different routes. This research has significant implications, particularly in public transport networks where the coverage of APC and AFC data is low or skewed. The developed models can assist in reducing the number of required sensors to monitor service utilisation more effectively and cost-efficiently.

The remainder of this paper is organised as follows. The existing literature is reviewed in "Literature review". "Methodology" specifies the problem and describes the model framework proposed in this paper. "Case study" elaborates on the setup of numerical experiments where the proposed methods are tested on three tram routes in Melbourne, Australia. Results are presented and discussed in "Results". In "Conclusion and future work", potential future improvements are discussed.

Literature review

The estimation of public transport passenger flows has received significant attention in transportation studies as it is the foundation of service planning and operations management. Traditionally, transit load estimation has been approached through a sequential four-step model (McNally 2007): (1) Generate trips in each zone based on land uses and demographic information; (2) Distribute trips to specific origins and destinations considering land use and network accessibility; (3) Select trips by public transport using choice models constructed by the analysis of user behaviour; and (4) Allocate trips to particular routes according to the network structure. However, the four-step model only constructs the relationships between origin–destination (OD) pairs during a specific time period of a regular day. While these models are useful for strategic planning, they fail to account for

the day-to-day and within-day variabilities in passenger demand and demand response to changes in service. This creates a data gap for operations management and in-detailed service planning analysis.

To address this gap, dynamic transit models have been developed that incorporate simulation techniques, such as BusMezzo (Cats 2011), FastTrips (Khani 2013), and schedule-based transit assignment methods offered by popular commercial tools like PTV Visum and Omnitrans. These models consider the passenger assignments as a sequence of travel decisions that are adapted based on traveller progress, enabling the representation of traveller responses to various conditions in the network. These models require a set of rules to represent passengers' behaviour and make assumptions on the interactions between demand and supply features in public transport (Gentile and Nökel 2016). However, some of these models require stated preference surveys that are specifically designed to estimate passenger behaviour, which can be labour-intensive and introduce biases and uncertainties to the results (Kagho et al. 2020). In addition, the computational cost associated with simulating a large number of agents with diverse behaviours poses a major challenge for dynamic transit models. While these models are suitable for strategic planning purposes, they are unable to provide real-time estimations of passenger loads for individual trips.

The predominant change in Intelligent Transportation Systems (ITS) has been catalysed by the quantity of data collected from various sources. Automated Passenger Counts (APC) systems provide accurate and automated recording of ridership rates and are increasingly adopted by transit operators. However, due to budget limitations, the common practice is to install APC systems on a subset of fleet vehicles (Strathman 1989; Siebert and Ellenberger 2020). Although this approach allows for obtaining the overall passenger flow distribution, it is insufficient for real-time monitoring of passenger loads for individual trips. In the absence of APC data for every service trip, some studies have employed supervised learning models using APC data as ground-truth labels (Moreira-Matias and Cats 2016; Jenelius 2019). Supervised learning models are capable of inferring the complex relationships between passenger flows and other independent variables, such as historical demand, dwell time, headway etc. These inferred relationships can then be utilised to estimate passenger loads in real time, even in the absence of direct measurement, given the aforementioned independent variables. However, one challenge with supervised learning models is that they only consider transit routes with APC systems. Transit routes without any APC systems lack historical observations and thus have to employ supervised learning models trained from other routes. This is typically a complex problem as the relationships between passenger flows and other independent variables may not be consistent across different routes. Unfortunately, no study to date has examined the application of supervised learning models to transit routes without any historical observations.

In contrast to APC systems, automatic fare collection (AFC) systems are often available across the entire public transport fleet for financial purposes. AFC systems typically record the time and location of boarding transactions, which are transferred from on-board (or stationary) fare validation devices to a data storage facility on a daily basis. Numerous attempts have been made in order to reconstruct passengers' trip chain using AFC data (Munizaga and Palma 2012; Gordon et al. 2013; Nassir et al. 2015, 2017, 2019). These trip chaining models are often fused with other transit data sources, such as automatic vehicle location (AVL) and general transit feed specification (GTFS) data, to infer the boarding and alighting locations of individual trajectories. Several studies have revealed the power of AFC data in constructing load profiles (Chu and Chapleau 2008; Luo et al. 2018) by utilising trip chaining models to monitor service utilisation. However, depending on the coverage and accuracy of AFC data, there may be trips that are not properly recorded or

cannot be inferred by trip chaining models. For example, in Chicago (Miller et al. 2018), only about 85% of passengers use fare cards for boarding, while others prefer paper tickets or cash options. In Santiago(Cantillo et al. 2022), Chile, fare evasion is a significant issue that raises serious concerns regarding operational cost recovery. In Melbourne, the free tram zone in the Central Business District is a blind spot in the fare collection system. Additionally, fare compliant passengers, such as those holding fare passes or those who validated their fare cards on a transfer/linked trip, may not be required to tap their card for a specific service trip. Furthermore, the trip chaining models rely on various behavioural assumptions to infer the alighting location, which may introduce additional inaccuracies. These are some potential sources of inaccuracy in AFC data that may vary across different agencies.

Some of the missing data can be addressed by scaling OD matrices using an expansion factor, assuming that the distribution of trips with unknown origins and destinations is similar to that of trips with inferred origins and destinations (Gordon et al. 2013). The expansion factor can be calculated according to the distribution of the part of transactions that are OD-known. However, this treatment cannot be applied to the case where the distribution of missing trips is not homogeneous. If there are significant differences in the spatial-temporal travel patterns, simple expansion methods might be biased and misleading (Gordillo 2006; Munizaga and Palma 2012). For example, in Melbourne's free tram zone where an entire sub-region is absent in the AFC data, scaling may not be an effective solution for data imputation.

In summary, it is challenging to derive unbiased measurements of passenger loads for individual trips from incomplete APC or AFC data. While some studies have demonstrated the potential value of using APC data to train supervised learning models, the transferability of such models across different transit routes remains to be determined. AFC data is often more obtainable than APC data for individual trips, but it is frequently negatively skewed in many cities. These issues motivated our research to improve existing models for public transport service load estimation using limited, incomplete, and skewed data sources.

Methodology

Problem specification

Given a set of service trips and stops, the main objective is to estimate the passenger load at a specific stop for a specific trip. A common approach is to reconstruct the trip chain of users using AFC data (Gordon et al. 2013; Munizaga and Palma 2012; Nassir et al. 2015, 2017, 2019) and aggregate the passenger flows to construct the load profiles for each trip (Luo et al. 2018). However, this approach may not produce accurate load profiles due to inherent limitations such as incomplete, inconsistent, or inaccurate records in the fare collection data (Kurauchi and Schmöcker 2017).

For example, AFC data from Melbourne Tram services is subject to various limitations that need to be addressed. According to a prior investigation, the touch-on rate for the month of June 2012 was 37%. This missing data can be attributed to several factors. Firstly, Melbourne tram services experience high fare evasion rates (Delbosc and Currie 2016), as passengers can board from any door without contact with the driver, leading to increased fare evasion. While ticket inspectors are employed to check valid tickets, they only board at

a few randomly selected stops, leaving the majority of trips uninspected. Secondly, while passengers are supposed to touch on when boarding, they are compliant if they hold a fare pass or have transferred from another service (train, tram, or bus). Another specific issue in Melbourne is the presence of a free tram zone located in the Central Business District (CBD) where passengers are actively discouraged to tap on or off tram services. If the proportion of missing trips were uniformly distributed across the service, routes, and trips, an expansion factor could be applied based on the distribution of transactions with identified origins and destinations (Gordon et al. 2013). For Melbourne, however, the distribution of missing information is likely to be non-uniform. Specifically, a large number of trips within the free tram zone are not recorded compared to other stops. The boundaries of the free tram zone are also critical, as some passengers may risk travelling a few extra stops without touching on, leading to a lower touch-on rate. Hence, a simple expansion factor may not be effective in accurately estimating passenger loads (Munizaga et al. 2020).

Automatic Passenger Counting (APC) systems were introduced to the Melbourne tram network in 2020; however, they are currently installed on only a subset of trams operating on three routes. It is still challenging to make it a dominant mode of data collection and obtain the precise passenger count for every vehicle due to budget limitations. Despite the limited coverage, the data collected from these APC devices can serve as reliable ground truth data to train machine learning models and discover correlations between passenger loads and other independent variables. The missing pieces of information can potentially be captured by other types of data collected for different purposes.

Considering the coverage of APC systems, we propose two supervised learning problems: (1) same-route estimation, and (2) cross-route estimation. Same-route estimation applies to routes where APC systems are installed on a subset of fleets. It utilises the passenger counts collected by APC systems on the same route as the ground truth labels. Cross-route estimation is designed for routes without any APC systems and employs passenger counts from routes as the ground-truth labels. The model is trained using data collected from routes with APC systems and then transferred to a different route that lacks historical observations in order to estimate the service load.

Section "Estimation framework" introduces the estimation frameworks that are applicable to both supervised learning problems mentioned earlier. Section "Feature development" provides a comprehensive overview of the input features extracted from exogenous sources of data, which are as inputs in the supervised learning model.

Estimation framework

This section proposes several solution methods for both same-route and cross-route estimation problems. Consider a set of service trips $\{1, 2, \dots, \tau \dots\}$ with a set of stops $\{1, 2, \dots, i \dots\}$ of transit route r in one direction. The main objective is to estimate the passenger load $l_{i,\tau}$ of trip τ at stop i. We assume that the following sets of information are available for a given trip:

1. A feature vector at stop i of trip τ derived from exogenous sources of data (See "Feature development"). Let $F^B_{i,\tau}$, $F^A_{i,\tau}$, and $F^L_{i,\tau}$ denote the feature vectors used for boarding flows, alighting flows, and passenger loads respectively.
2. Passenger OD flow $X_{i,j,\tau}$ from stop i to stop j of trip τ derived from the incomplete AFC data

Benchmarks

A straightforward approach is to feed features into a supervised learning model and estimate passenger loads directly. Let M^L denote the supervised learning model trained with the ground-truth label of passenger loads. The estimated passenger load $l_{i,\tau}$ at stop i of trip τ is given by:

$$\hat{l_{i,\tau}} = M^L(F^L_{i,\tau}) \tag{1}$$

Passenger loads can also be estimated indirectly by modelling boarding and alighting flows. Let M^B and M^A denote the supervised learning model trained with the ground-truth label of boarding flows and alighting flows, respectively. The estimated boarding flows $\hat{b_{i,\tau}}$ and alighting flows $\hat{a_{i,\tau}}$ at stop i of trip τ are given by:

$$\hat{b_{i,\tau}} = M^B(F^B_{i,\tau}) \tag{2}$$

$$\hat{a_{i,\tau}} = M^A(F^A_{i,\tau}) \tag{3}$$

Finally, the estimated passenger loads $\hat{l_{i,\tau}}$ is calculated as a function of boarding and alighting flows 4:

$$\hat{l_{i,\tau}} = \sum_{k=1}^{k=i} (\hat{b_{k,\tau}} - \hat{a_{k,\tau}}) \tag{4}$$

However, these two estimation methods overlook the relationships between boarding flows, alighting flows, and passenger loads on-board. Many researchers believe these relationships capture the long-range dependencies in passenger flow data (Li and Cassidy 2007; McCord et al. 2010; Sun et al. 2021; Cheng et al. 2021). For example, the number of alighting passengers at each stop is highly dependent on the passenger loads on-board, as the alighting passengers must have boarded the vehicle at previous stops. In order to capture these relationships, we propose three different estimation methods: a two-stage estimation, an OD-based estimation, and a segment-based estimation.

Two-stage load estimation

The two-stage method captures the relationships between passenger loads on-board and the alighting flows. It involves a coarse estimation and a calibration step using the estimated passenger loads on-board. In the calibration step, the estimated passenger loads are included as extra information (added feature) to update the estimation of alighting flows.

First, we train separate models for boarding M^B and alighting M^A using Eqs. 2 - 4 to obtain a coarse estimation of the load for both training and test data. This provides us with a coarse estimation of passenger loads on board $\hat{l_{i,\tau}}$ for each stop i of trip τ.

To model the correlation between alighting flows and passenger loads on-board, we incorporate the passenger load at the previous stop $l^p_{i,\tau}$ into the feature space of the alighting model. The passenger load at the previous stop is calculated using Eq. 5.

$$l^p_{i,\tau} = \begin{cases} 0 & i = 1 \\ \hat{l_{i-1,\tau}} & i > 1 \end{cases} \tag{5}$$

We then train a calibration model N^A with $F_{i,\tau}$ and $l^p_{i,\tau}$ to refine the estimation of alighting flows. The alighting demand $\hat{\alpha_{i,\tau}}$ at stop i of trip τ is then updated by N^A, incorporating the feature vector and passenger loads at the previous stop.

$$\hat{\alpha_{i,\tau}} = N^A(F^A_{i,\tau}, l^p_{i,\tau}) \tag{6}$$

Finally, the final estimation of passenger load ($\hat{\gamma_{i,\tau}}$) of trip τ at stop i is calculated using the equation:

$$\hat{\gamma_{i,\tau}} = \sum_{k=1}^{k=i} (\hat{b_{k,\tau}} - \hat{\alpha_{k,\tau}}) \tag{7}$$

OD-based load estimation

The relationship between passenger flows and input features may not be consistent over different routes, which can introduce biases in cross-route estimation when the model is trained with the ground-truth labels from other routes. On the other hand, AFC data, despite having incomplete, inconsistent, and inaccurate records, is available for all service trips and may preserve the unique characteristics of the specific route.

The OD-Based estimation integrates the supervised learning model with demand patterns from AFC data. Instead of estimating boarding and alighting volumes independently, the OD-based model uses both supervised learning models and OD flow matrix retrieved from AFC data. For AFC data consisting of both tap-on and tap-out records, the OD matrix can be directly derived by aggregating individual trajectories. In cities with tap-on only data, the OD matrix can be derived through trip chaining models (Nassir et al. 2011; Munizaga and Palma 2012; Nassir et al. 2015, 2017, 2019). Alternatively, the OD matrix can be derived from other data sources, such as human survey or strategic model. While the static OD matrix may not reflect the passenger flow at the trip level, it is used to generate boarding and alighting probabilities for every stop in this method. The OD-based estimation can be further classified into boarding-based estimation and alighting-based estimation.

Boarding-based Load Estimation The boarding-based estimation method estimates boarding flows using supervised learning models and calculates the alighting demand from boarding demand using the OD flow matrix retrieved from AFC data. First, we train the boarding model M_B and apply Eq. 2 to predict the boarding flows $\hat{b_{j,\tau}}$ at stop j of trip τ.

Then, the alighting probability $P_{j,i}$ is constructed from the AFC OD probabilities by aggregating passenger counts of individual service trips. $P_{j,i}$ represents the probability that a passenger boarding at stop j will alight at atop i. $X_{j,i}$ denotes the passenger flows from stop j to stop i recorded by AFC data. The alighting probability is calculated as follows:

$$P_{j,i} = \frac{\sum_{\forall \tau} X_{j,i,\tau}}{\sum_{\forall k>j} \sum_{\forall \tau} X_{j,k,\tau}} \tag{8}$$

After getting the estimation of boarding flows $\hat{b_{j,\tau}}$ of trip τ at stop j and the alighting probability $P_{j,i}$, the alighting flows can be estimated directly from boarding volumes and OD distributions. It is assumed that trips on the same route share the same alighting probability. The alighting flow at stop i of trip τ is estimated by:

$$\hat{a_{i,\tau}} = \sum_{j=1}^{j=i-1} \hat{b_{j,\tau}} P_{j,i} \tag{9}$$

With the estimation of boarding and alighting flows at each stop, the load $\hat{l_{i,\tau}}$ of trip τ at stop i is calculated by Eq. 4.

Alighting-based Estimation The alighting-based estimation uses the supervised learning model to estimate alighting flows $\hat{a_{j,\tau}}$ at stop j of trip τ. The boarding flows are then estimated using the boarding probability. Similar to the alighting probability calculated by Eq. 8, the boarding probability $Q_{i,j}$ from stop i to stop j can also be derived from AFC data, representing the probability that a passenger who alights at stop j originated from stop i:

$$Q_{i,j} = \frac{\sum_{\forall \tau} X_{i,j,\tau}}{\sum_{\forall k<j} \sum_{\forall \tau} X_{k,j,\tau}} \tag{10}$$

Then, the boarding flow at stop i of trip τ is calculated as follows:

$$\hat{b_{i,\tau}} = \sum_{j=i+1}^{j=I} \hat{a_{j,\tau}} Q_{i,j} \tag{11}$$

With the estimation of boarding and alighting flows at each stop, the load $\hat{l_{i,\tau}}$ of trip τ at stop i is calculated by Eq. 4.

Figure 1 illustrates the structure of the (a) indirect estimation (b) two-stage estimation, (c) boarding-based estimation, and (d) alighting-based estimation. The bold line indicates the flow of information used for estimating passenger loads. Compared to indirect estimation, all other methods incorporate additional information to model the

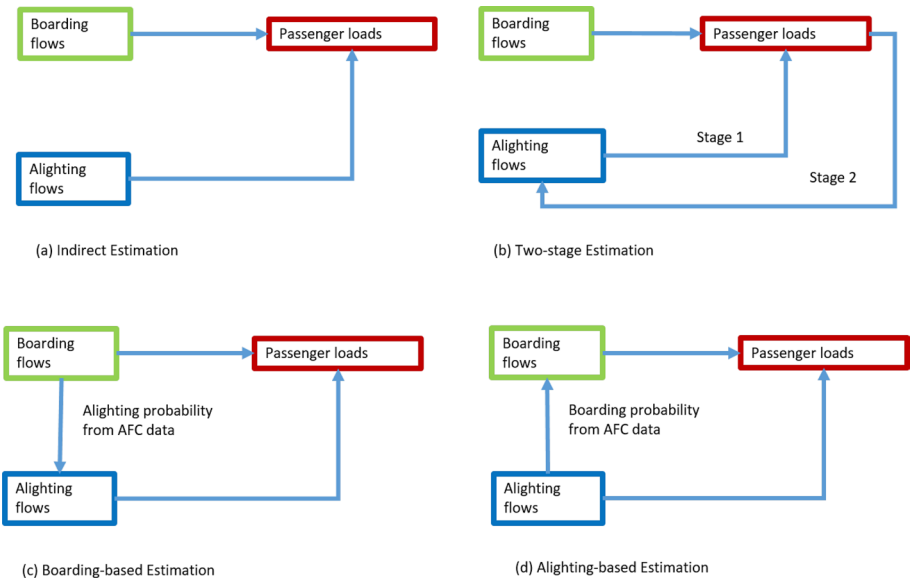

Fig. 1 Structure of the proposed estimation framework: **a** Indirect estimation model; **b** Two-stage estimation model; **c** Boarding-based estimation model; (d) Alighting-based estimation model

dependencies between boarding flows, alighting flows, and passenger loads. The two-stage estimation method utilises the supervised learning model to capture the relationship between passenger loads on-board and alighting flows. The two OD-based methods, on the other hand, connect boarding flows and alighting flows through the OD distribution probabilities derived from AFC data.

Segment-based estimation

Both the two-stage estimation and the OD-based estimation methods involve estimating boarding and alighting flows, followed by the calculation of passenger loads. However, there is a potential risk when the estimation of boarding or alighting flows is not accurate at a particular stop. These errors can propagate downstream, resulting in biased estimated loads, even if the estimates for the downstream stops are perfect. To mitigate this cascading failure effect, a segment-based estimation method is proposed, where stops in different segments are estimated using hybrid methods.

We start by training the boarding model M_B and the alighting model M_A using the feature vector and ground-truth labels. This allows us to obtain the estimation of boarding flows $\hat{b_{i,\tau}}$ and alighting flows $\hat{a_{i,\tau}}$ using Eqs. 2 and 3

Next, the stops of a route are split into three segments: the peak segment (PS), inbound segment (IS), and outbound segment (OS). The peak segment typically contains major destinations in the study period, such as the Central Business District (CBD), where most of the financial, legal administrative, and retail facilities are concentrated, attracting a diverse range of workers. Stops located upstream of the peak segment are classified as the inbound segment, as most trips generated in this region are likely to terminate in the peak segment. Stops located downstream of the peak segment are referred to as the outbound segment, as they primarily serve trips originating from the peak segment.

In the inbound segment, most of those trips end up at common destinations in the CBD. As a result, trips across different routes are generated in a similar manner, and boarding flows across different routes may be most affected by trip production features. For example, the number of trips generated in the inbound segment is likely to be proportional to the population in the areas within the inbound segment. Hence, we employ the supervised learning model to estimate the boarding flows by Eq. 2. Regarding alighting flows, trips that terminate the inbound segment do not share common destinations. The relationships between alighting flows and other independent variables are likely to vary across different routes. Consequently, the alighting flows in the inbound segment are estimated by the boarding-based method outlined in Eqs. 8 - 9. This method relies solely on the demand patterns from AFC, which capture the unique characteristics of a specific route.

In contrast, in the outbound segment, trips originating from the peak segment are more likely to exhibit a similar attraction pattern. A similar correlation between alighting flows and other independent variables across different routes is expected. For stops in the outbound segment, we employ the alighting-based method outlined in Eqs. 8 - 9. Alighting flows are calculated using supervised learning models, while boarding flows are determined by the boarding probability.

Hence, the boarding and alighting flows at stops in the inbound and outbound segments are calculated by the following equations. Let $\beta_{i,\tau}$ and $\alpha_{i,\tau}$ denote the final estimation of boarding and alighting flows at stop i for trip τ, respectively.

$$\hat{\beta_{i,\tau}} = \begin{cases} \hat{b_{i,\tau}}, & \text{if } i \in IS \\ \sum_{j=i+1}^{j=I} \hat{a_{j,\tau}} Q_{i,j}, & \text{if } i \in OS \end{cases} \tag{12}$$

$$\widehat{\alpha_{i,\tau}} = \begin{cases} \sum_{j=1}^{j=i-1} \widehat{b_{j,\tau}} P_{j,i}, & \text{if } i \in IS \\ \widehat{a_{i,\tau}}, & \text{if } i \in OS \end{cases} \tag{13}$$

To prevent load estimation errors from propagating along the load profile, we calculate the load from the first stop for stops in the inbound segment; while for stops in the outbound segment, we calculate the load from the last stop. The load at the first and last stop is always zero, and the estimation errors will only cascade within the respective segment.

Estimating loads using the OD-based method for stops in the peak segment can lead to significant errors due to high volumes and variations in passenger flows. However, as this segment is typically served by multiple lines, passenger flows in this region may exhibit a similar spatial distribution across different routes. Although passengers travelling to this segment come from various routes and directions in the network, they ultimately end up at common destinations in the CBD. To address this, we utilised another supervised learning model to directly estimate the load, taking into account the coordinates of the stops within the CBD and the loads estimated by the boarding-based model at the end of the inbound segment. The former captures the effect of land use, as adjacent stops are likely to have similar land use types, while the latter captures the variation over service trips. In this way, although errors are expected in the peak segment due to high variation, they will not cascade to stops in the inbound and outbound segments.

In summary, the load is given by Eq. 14

$$\hat{l_{i,\tau,r}} = \begin{cases} \sum_{k=1}^{k=i} (\hat{\beta_{k,\tau}} - \hat{\alpha_{k,\tau}}), & \text{if } i \in IS \\ \sum_{k=i}^{k=I} (\hat{\alpha_{k,\tau}} - \hat{\beta_{k,\tau}}), & \text{if } i \in OS \\ K^L(G_i, \hat{l_{\delta,\tau}}), & \text{if } i \in PS \end{cases} \tag{14}$$

where K^L is the supervised learning model trained by data collected in the peak segment, G_i is the coordinates of stop i, and $\hat{l_{\delta,\tau}}$ is the load estimated by the boarding-based method at the last stop δ in the inbound segment for trip τ.

Figure 2 shows the model framework of segment-based method.

Feature development

This section discusses the input features used in the supervised learning models. Most of the transportation models categorise variables into travel demand features and service supply features (Gentile and Nökel 2016).

Demand features

Travel demand features consist of time-based features and census information that reflect the spatial-temporal pattern of demand distribution.

Time-based Features Time-based features include the day of the week and time of day, which capture the variations in passenger flow over time.

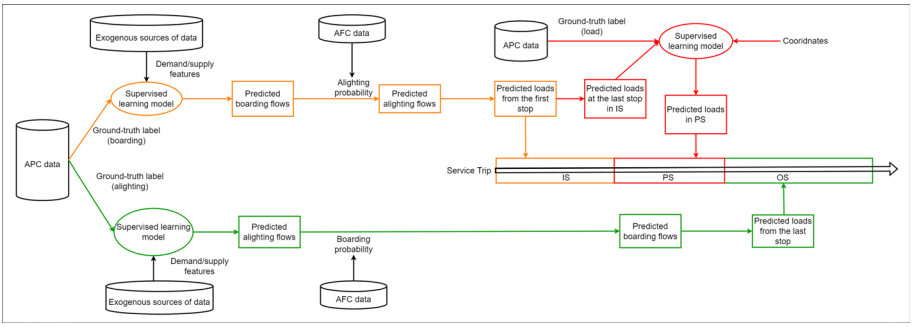

Fig. 2 Framework of the segment-based model

Census Data Census data refers to surveys conducted on a given population. Some information, such as population and vehicle ownership, has been widely used in traffic demand modelling (McNally 2007; Cats 2011) to generate trips in each zone. Census data is typically aggregated by geographic area and provides insights into demand distribution across space. For stop i, we select all the geographic areas that are within walking distance and represent the set of areas by A_i. In this study, the census attributes include residential population and vehicle ownership.

Let p_a and v_a denote the population and household vehicle ownership in geographic area a. The stop-level population P_i and household vehicle ownership V_i at stop i are given by:

$$P_i = \sum_{\forall a \in A_i} p_a \tag{15}$$

$$V_i = \sum_{\forall a \in A_i} v_a \tag{16}$$

Supply features

Service supply features reflect the passenger allocation due to service performance, including dwell times and headway. These features can be obtained from (Automatic Vehicle Location) AVL data, which captures the variability of load distribution between vehicles.

Dwell Time Dwell time is the length of time that a vehicle spends at a stop. Dwell time has a strong correlation with the total number of boarding and alighting passengers (Sun et al. 2014; Glick and Figliozzi 2017), and some studies use dwell time (Bie et al. 2015; Sun et al. 2021) as a proxy of passenger demand. The dwell time $D_{i,\tau}$ of trip τ at stop i is calculated as follows:

$$D_{i,\tau} = DT_{i,\tau} - AT_{i,\tau} \tag{17}$$

where $DT_{i,\tau}$ represents the actual departure time and $AT_{i,\tau}$ represents the actual arrival time at stop i of trip τ.

Headway Headway refers to the time between two successive service arrivals at a specific stop. The total passenger demand at a stop is allocated to vehicles based on their arrivals (Han and Wilson 1982). Studies assuming a random (Poisson) process for passengers

to come to stops commonly deduct that the number of passengers boarding the service at a given stop is proportional to the headway (Turnquist 1978; Jenelius 2019). The headway $H_{i,\tau}$ at stop i of trip τ is calculated as follows:

$$H_{i,\tau} = DT_{i,\tau} - AT_{i,\tau-1} \tag{18}$$

where $DT_{i,\tau}$ represents the actual departure time of trip τ and $AT_{i,\tau-1}$ represents the actual arrival time of the preceding trip at the arrival of stop i.

XGBoost model

Alternative supervised learning models can be used for models M^B, M^A, M^L, N^A, and K_L in the proposed model frameworks proposed in "Feature development". In this research, Extreme Gradient Boosting (XGBoost) model is chosen due to its proven effectiveness in various domains with high accuracy and relatively low computational time.

XGBoost is a tree-based model that assumes that the complex interactions within the data can be represented by a tree (Chen and Guestrin 2016). Tree-based models divide the feature space into a series of rectangles, where each rectangle corresponds to a simple model, often just a constant. XGBoost model was specifically developed to enhance the performance of decision tree models by combining multiple weak predictors, resulting in a highly accurate model and more precise predictions. For a given dataset with m features, the XGBoost model employs K additive trees to create the ensemble model:

$$\hat{y} = \sum_{i=0}^{K} \hat{f}_n(x), \hat{f}_n(x) \in F \tag{19}$$

where $F = \{w_{q(x)}\}(q : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ is the space of regression trees, q represents the tree structure that maps the input to the corresponding leaf index, T is the number of leaves in the tree. Each f_k corresponds to an independent tree structure q and leaf weights w. The tree structure uses the decision rules in the tree (given by q) to classify features into leaves and calculate the final prediction by summing up the weights in the corresponding leave (given by w). The model is trained by minimising the following loss function:

$$\begin{aligned} U(\phi) &= \sum_n u(\hat{y_n}, y_n) + \sum_k \Omega(f_k) \\ \text{where } \Omega(f_k) &= \gamma T + \frac{1}{2}\gamma \|w\| \end{aligned} \tag{20}$$

where u is a differentiable convex loss function that measures the difference between the prediction $\hat{y_n}$ and the target y_n, Ω is the regularisation term that penalises the complexity of the regression tree.

Case study

APC data and AFC data

The proposed methods were tested on three tram routes (Route 11, Route 86, and Route 96) in Melbourne, Australia. For this experiment, only trips departing from the first stop during

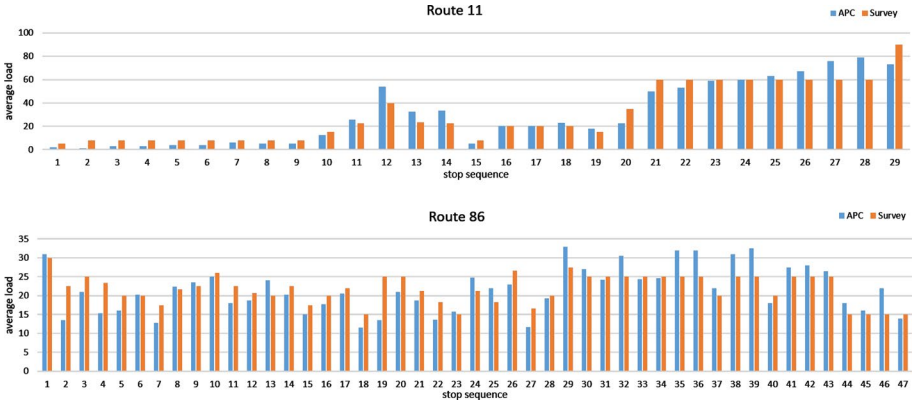

Fig. 3 The average passenger loads recorded by APC and survey data at inspected stops for Route 11 and Route 86

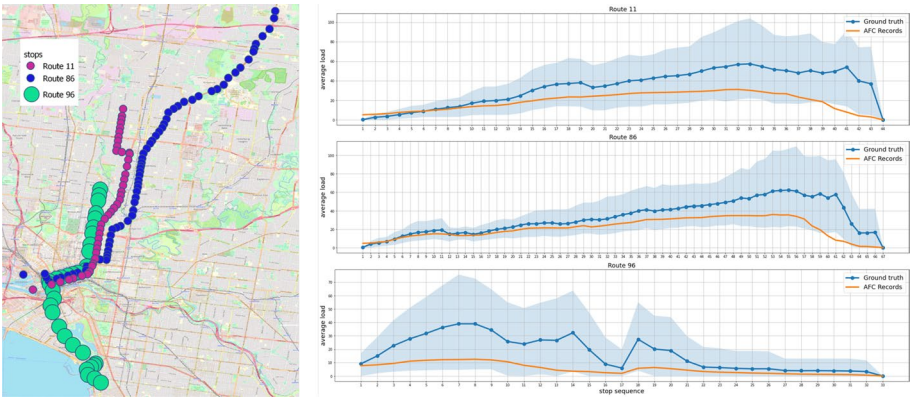

Fig. 4 (left) Geographic distribution of route 11, 86, and 96 in Melbourne, Australia; (right) Average load at each stop recorded by APC (truth) and AFC data

the morning peak (7:00-10:00) on weekdays from, 01/02/2020 to 16/03/2020, in the direction towards the Central Business District (CBD) were considered. Figure 4 (left) shows the geographic distribution of these three tram lines.

A subset of the fleet of these three routes was equipped with APC systems. The passenger flows recorded by the APC systems were used as the ground truth data. To validate the use of APC as ground truth data, a comparison was made between the ridership data from APC systems and the ridership recorded by authorised officers at certain stops. The authorised officers would board the tram at randomly selected points on the network and remain on-board for a few stops, collecting data on checked tickets and estimating the number of on-board passengers. During the study period, only a limited number of vehicles with APC devices were at some stops. We found 8 trips for Route 11 and 14 trips for Route 86 from APC data that can be matched with the on-board survey. Most of the stops on Route 96 were not inspected, resulting in no matched trips. Figure 3 shows the average load from APC and survey data at the inspected stops of the matched trips. The passenger loads

recorded by both datasets show a rough consistency, with an average deviation of 6.18 passengers for Route 11 and 7.19 passengers for Route 86. The deviation is close to 10% of the seating capacity of the vehicles studied (64 passengers), indicating that the measurement errors are not significant in terms of making operational decisions and providing real-time crowding information. Based on these findings, it can be concluded that the raw APC data is consistent with real observations and can be used as the ground truth data. After data cleaning and pre-processing, a total of 161 trips with 44 stops were obtained for Route 11, 147 trips with 67 stops for Route 86, and 292 trips with 33 stops for Route 96.

Myki is the AFC system used for the electronic payment of fares in Melbourne, Australia. It records individual fare validation trajectories, which allow for the estimation of passenger flow volumes between identified origin and destination pairs through trip chaining (Nassir et al. 2011, 2015, 2017). In this study, we employ the rule-based trip chaining procedure proposed by these studies to process Myki data, which includes the inference of boarding and alighting stops. For the study period, 77.9% of the Myki transaction locations have been successfully inferred. By aggregating these transactions of each service trip, the passenger loads can be derived. To account for the undermined transactions, an expansion factor of 21% is calculated. This expansion factor represents the ratio between the total number of determined transactions and the total number of undetermined transactions. The passenger loads are scaled via the expansion factor calculated for each trip made.

Figure 4 (right) represents a comparison of the average passenger loads recorded by APC data and AFC data at each stop during the study period. The shaded area represents the variations observed in the APC data, with 10^{th} percentile and 90^{th} percentile of passenger loads. The x-axis represents the stop sequence and the y-axis represents the number of passengers on-board. Despite successfully identifying most transaction locations in the AFC data and scaling the loads using the expansion factor, we still observe significant underestimations in the records from the AFC data at each stop. As previously discussed, this discrepancy arises from unrecorded trips due to fare evasion or the free tram zone. The distribution of this missing information is non-uniform, so a simple scaling factor is inadequate. This motivates the introduction of supervised learning models to improve the estimation of passenger loads.

Exogenous sources of data

For the supervised learning models, the following input features are extracted from the exogenous data sources discussed in "Feature development" and then matched with the APC data.

Census data, obtained from the Australian Bureau of Statistics, is aggregated based on statistical area defined by the Australian Statistical Geography Standard. The relevant census datasets for estimating passenger loads on trams include residential population and vehicle ownership. These features are calculated using Eqs. 15 - 16, with a chosen walkable distance (from zone centroid to stop) of 360 m, which represents the median length of a walking trip to a tram stop in Melbourne (Eady and Burtt 2019).

Automated Vehicle Location (AVL) data tracks the location of each vehicle at specific time points along the route. The General Transit Feed Specification (GTFS) data provides the schedule information, which is essential for matching the trips among timetables. From the GTFS data and AVL data, the following features are derived: day of the week, arrival time, and service supply features including dwell time and headway.

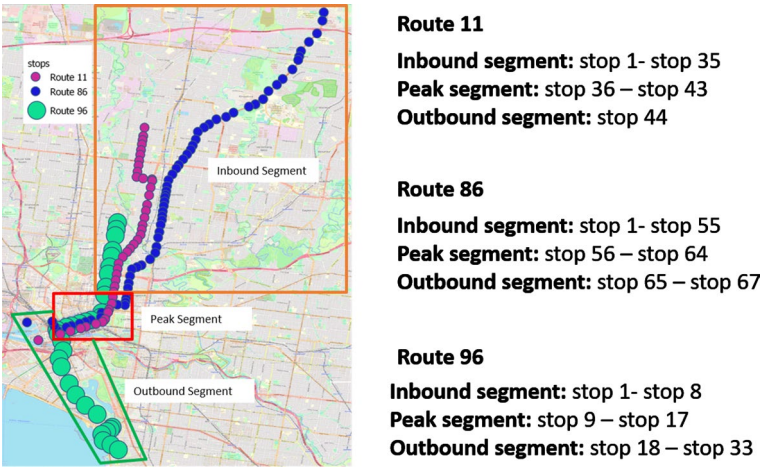

Fig. 5 Stop segmentation used in the segment-based model

Table 1 Number of samples in the training and test set in this study

	Route 11	Route 86	Route 96
Same-route estimation			
Training set	4928	7303	4554
Test set	2156	2546	5082
Cross-route estimation			
Training set	19485	16720	16933
Test set	2156	2546	5082

Model fitting

For the same-route estimation problem, data from the first two weeks (03/02/2020 - 28/02/2020) are used as the training set to derive the passenger flow estimation model (M^A, M^B, M^L, N^A, and K^L). Data from the last two weeks are used as the test set (02/03/2020-13/03/2020).

For the cross-route estimation problem, data from the other two routes are used as the training set. For instance, to estimate the service loads on Route 96, the passenger flow estimation models are trained with the APC data from Route 11 and Route 86. The same test set is still utilised as in the same-route estimation problem.

Each stop of every service trip is transformed into a sample. The number of samples used for both same- and cross-route estimation is presented in Table 1. A tenfold cross-validation approach is employed to tune the model. The parameters to be determined include tree depth, number of trees, and the subsample ratio of columns during tree constructing. A grid is used to choose the optimal parameters.

For the segment-based model (See "Estimation framework"), the stops of each route need to be divided into three segments: the inbound segment, peak segment, and outbound segment. Figure 5 shows the stop segmentation. The inbound segment includes tram stops in the north suburbs, which are major residential areas in Melbourne. During the morning

peak, it is expected that many commuting trips originate from this region. The peak segment corresponds to the Melbourne central business district (CBD), where the majority of financial, legal administrative, and retail facilities are located. It attracts a diverse range of workers during the morning peak. In Melbourne, it's worth noting that the peak segment overlaps with the free service zone, where passengers are not required to touch-on, resulting in most boarding flows not being recorded by AFC data. The outbound segment consists of Dockland, South Melbourne, Albert Park, and St Kilda. These suburbs are known for their recreational facilities, such as harbours, parks, and beaches. While they may attract some trips, they are not as popular as the CBD during weekday mornings.

Performance evaluation

The estimation results are evaluated using the mean absolute error (MAE) and root mean square error (RMSE). The MAE and RMSE of N samples are calculated as follows:

$$MAE = \frac{1}{N} \sum_{n} |\hat{l_n} - l_n| \tag{21}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{n} (\hat{l_n} - l_n)^2} \tag{22}$$

where $\hat{l_n}$ and l_n represent the estimated and actual passenger loads of sample n, respectively. The RMSE metric assigns a relatively higher weight to large errors compared to the MAE metric.

Results

Using the same input features and test data (02/03/2020 - 13/03/2020) during the morning peak period in the inbound direction, five alternative models are compared, and the results are summarised in this section.

- **AFC load records**: Records in the AFC data after trip chaining (See "APC data and AFC data").
- **Direct load estimation method**: The benchmark method that estimates passenger loads directly using supervised learning (See "Estimation framework").
- **Indirect load estimation method**: The benchmark method that estimates boarding and alighting flows independently using supervised learning, and then calculates loads (See "Estimation framework").
- **Two-stage load estimation method**: The proposed two-stage estimation method, in which the estimation of passenger loads is then used to update the estimation of alighting flows (See "Estimation framework").
- **Boarding-based load estimation method**: The proposed boarding-based estimation method, in which the boarding flows are estimated by the supervised learning model and alighting flows are calculated by the alighting probability obtained from AFC data (See "Estimation framework").
- **Alighting-based load estimation method**: The proposed alighting-based estimation method, in which the alighting flows are estimated by the supervised learning model

and boarding flows are calculated by the boarding probability obtained from AFC data (See "Estimation framework").

- **Segment-based load estimation method**: The proposed segment-based estimation method, in which stops are split into segments and the passenger loads in each segment are estimated by separate methods (See "Estimation framework").

Table 2 presents the accuracy of estimation based on absolute value, measured by mean absolute error (MAE) and root mean squared error (RMSE). The performance can also be visualised by the predicted average loads at each stop. Figure 6 shows the average estimated load profiles of the same-route estimation problem, and Fig. 7 shows the average estimated load profiles of the cross-route estimation problem.

For the same-route estimation problem, the results show that the simple estimation from AFC records is significantly improved by introducing the supervised learning models. For Route 11 and Route 86, the two-stage estimation and OD-based estimation outperform the direct and indirect estimation because both of these methods link the alighting flows to the boarding flows. However, for Route 96, the errors of OD-based estimation, whether boarding-based or alighting-based, are considerably high. Route 96 serves both CBD commuters and passengers travelling from CBD to downstream stops, and since Melbourne CBD is a free tram zone, many trips starting from CBD are not recorded in the AFC data. Therefore, the alighting probability obtained from AFC data for Route 96 does not include passengers travelling from CBD to downstream stops, leading to significant errors at stops within the CBD. These errors cascade to all downstream stops and affect the estimations. The segment-based method addresses this issue for Route 96 by splitting stops into segments, containing errors within each segment rather than cascading to all downstream stops. In summary, for same-route estimation, the present findings suggest that supervised learning models should be used in public transport networks where the AFC data is screwed.

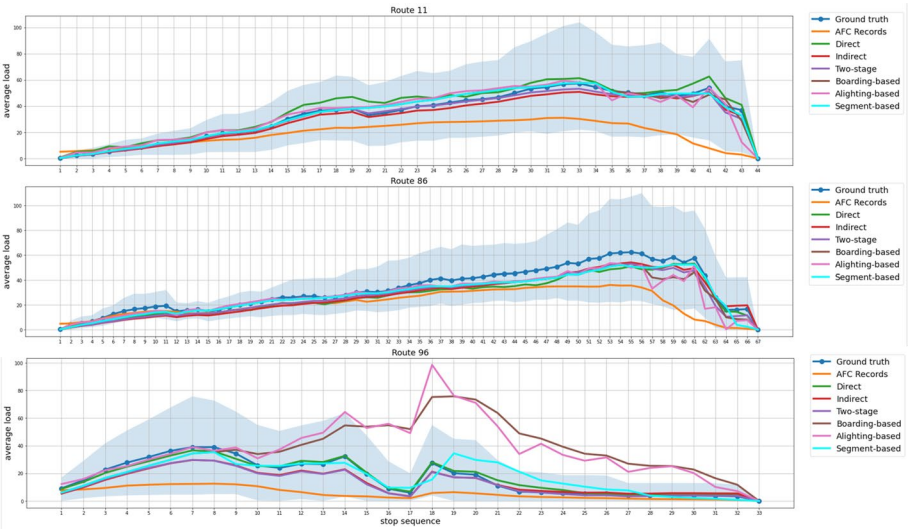

Fig. 6 The average passenger loads estimated by three methods for route 11, 86, and 96 (same-route estimation)

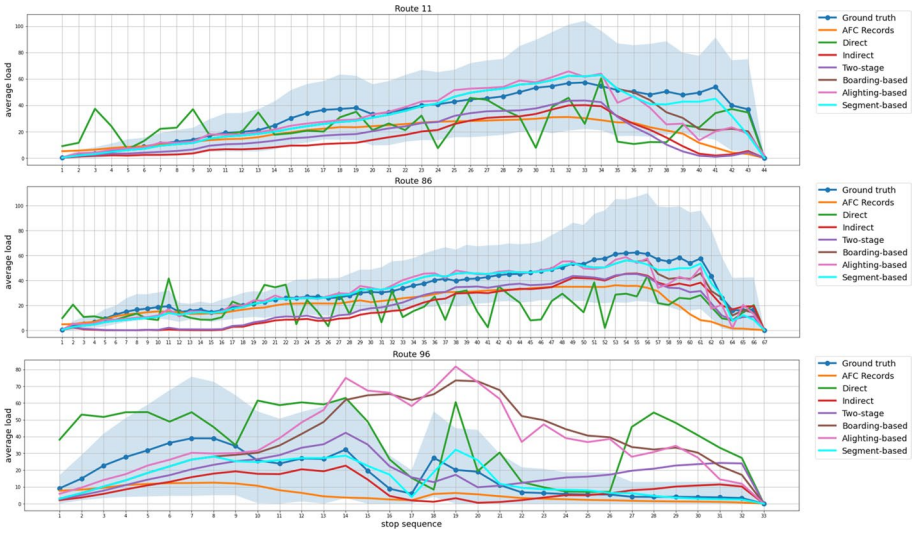

Fig. 7 The average passenger loads estimated by three methods for route 11, 86, and 96 (cross-route estimation)

Table 2 Performance comparison between the proposed method and benchmarks in terms of absolute value

Method	Route 11		Route 86		Route 96	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Same-route estimation						
AFC records	19.8	30.01	13.9	23.02	15	22.86
Direct estimation	15.7	21.57	11.9	17.58	11.7	**15.99**
Indirect estimation	12.5	18.72	13.1	19.67	11.6	16.72
Two-stage estimation	**12.13**	18.57	12.79	19.51	**11.31**	16.46
Boarding-based estimation	12.19	**18.38**	**10.66**	**16.01**	25.23	32.92
Alighting-based estimation	12.64	19.03	10.95	16.57	24.37	32.68
Segment-based estimation	12.49	18.8	10.73	16.05	11.99	16.65
Cross-route estimation						
AFC records	19.8	30.01	13.9	23.03	15	22.86
Direct estimation	19.9	28.03	19.6	27.39	25.4	32.88
Indirect estimation	20.2	29.32	16.8	22.03	13.5	19.41
Two-stage estimation	18.2	28.2	15.48	21.62	16.22	21.41
Boarding-based estimation	12.93	20.42	12.06	16.93	28.77	36.61
Alighting-based estimation	12.89	21.39	12.63	17.78	27.83	36.02
Segment-based estimation	**12.68**	**19.53**	**11.99**	**16.75**	**11.87**	**16.4**

The bold numbers highlight the minimum MAE and RMSE, indicating the best performance achieved by the respective models

For the cross-route estimation problem, we observe that the MAE and RMSE of the direct estimation, indirect estimation, and two-stage estimation are significantly increased. These methods rely on supervised learning models trained from entirely different routes. There is a risk that the correlation between passenger flows and the input features is not consistent across different routes. This inconsistency introduces bias in the estimations using pure supervised learning models. We also discover that the direct estimation method gives the highest error. As shown in Fig. 7, a random fluctuation is observed in the average load estimated by the direct method, without following any patterns. This indicates that the correlation between passenger loads and input features is entirely different across tram routes. On the other hand, the indirect estimation method slightly outperforms the direct method. This suggests that the correlation between boarding/alighting flows and extracted features, such as headway and population, is more consistent between different routes.

Nonetheless, the performance of the OD-based estimation method demonstrates a more robust performance even if the training data is from different routes. This method only uses supervised learning models to estimate either boarding or alighting flows and calculate the remaining flows based on the OD flow matrix obtained from AFC data. It relies not only on the APC data from other routes but also on the information from the AFC data of the same route. Therefore, if the correlation between features and passenger flows is inconsistent across different routes, it will have a lesser impact on the estimation performance. The OD-based method sorely relies on aggregated AFC data to derive the alighting probability, which means that missing counts from a single service will not have a significant impact on the estimation of passenger flows. However, in cases where the AFC data is significantly biased, such as Route 96, the performance of the OD-based model remains poor.

The segment-based method outperforms all other methods for the cross-route estimation problem. This model is developed based on the assumption that only certain parts

Fig. 8 The average passenger boarding flows for Route 11, Route 86, and Route 96 (cross-route estimation) are estimated by two methods. The stops are divided into segments: inbound segment (orange), peak segment (red), and outbound segment (green). The pure supervised learning model performs better in the inbound segment but worse in the outbound segment, compared to the OD-based method. (Color figure online)

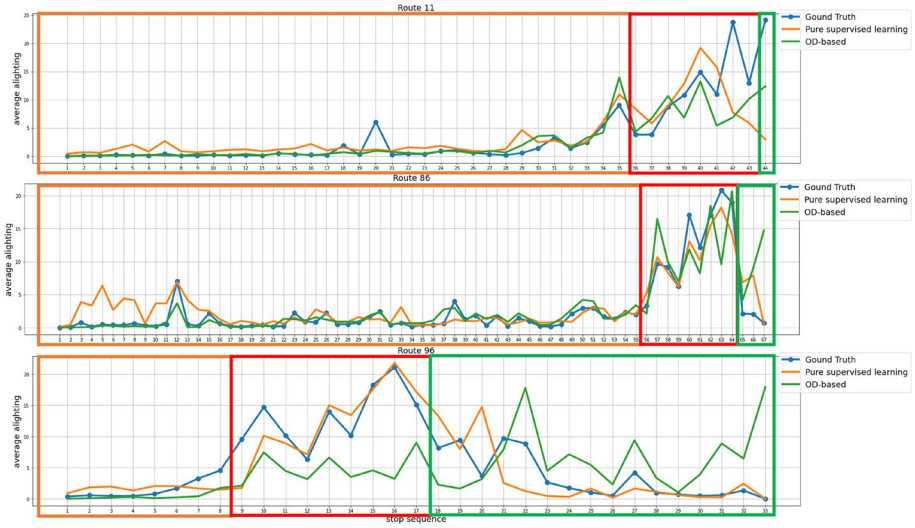

Fig. 9 The average passenger alighting flows for Route 11, Route 86, and Route 96 (cross-route estimation) are estimated by two methods. The stops are divided into segments: inbound segment (orange), peak segment (red), and outbound segment (green). The pure supervised learning model performs better in the outbound segment but worse in the inbound segment, compared to the OD-based method. (Color figure online)

of the relationships between passenger flow data and input features remain consistent across different routes, while they may vary among traffic analysis zones with different functionalities. Therefore, supervised learning models can only be effectively applied to those segments that exhibit a similar correlation between passenger flows and other variables. This hypothesis is supported by Figs. 8 and 9, which compare the boarding and alighting flows estimated by two methods: (1) pure supervised learning model (Eq. 2 - Eq. 3), where boarding or alighting flows are estimated sorely by supervised learning models and (2) OD-based model (Eqs. 8 - 11), where boarding or alighting probabilities from AFC data are utilised in addition to the estimation by supervised learning models. Stops are divided into inbound segment (orange), peak segment (red), and outbound segment (green).

For the boarding flows shown in Fig. 8, we observe that the performance of the pure supervised learning method is better than the OD-based method for all three routes. This can be attributed to the converging natural topology of the network, which consists of radial lines. As most trips generated in this region will terminate in the peak segment, the relationships between boarding flows and other independent variables are likely to be consistent across routes. Therefore, the pure supervised learning models are effective in estimating boarding flows in the inbound segments. On the other hand, for trips generated in the outbound segment, which do not have common destinations across different routes, the estimations from the pure supervised learning method are more likely to be biased. For instance, as shown in Fig. 8, for Route 96, the pure supervised learning method consistently overestimates the boarding flows in the outbound segment from stop 19 to stop 33. In contrast, the OD-based method yields better results, making it the preferred choice in the inbound segment.

In contrast, for alighting flows in Fig. 9, we observe that the pure supervised learning method does not perform well in the inbound segment. This is because trips that terminate

in this segment do not have similar destinations across different routes. As a result, the OD-based method is used to estimate alighting flows at stops in the inbound direction, and it clearly yields better results. On the other hand, in the outbound segment, the performance of the supervised learning method is better than the OD-based method. In this region, many trips may be attracted in a similar way, resulting in consistent relationships between alighting flows and attraction features across different routes. Hence, the pure supervised learning model is utilised to estimate alighting flows in the outbound segment.

In the peak segment, which is the major destination for many commuting trips during the morning peak, the variation in passenger flows can be primarily attributed to the land use in the CBD. Passenger loads in this region are estimated directly using the input feature of coordinates, which captures the land use information, and the estimated load at the end of the inbound segment, which captures the passenger flow variations in the inbound segment. Hence, as shown in Fig. 7, the segment-based method consistently outperforms all other methods for stops in the peak segment (stop 36-43 for Route 11; stop 56 to 64 for Route 86; and stop 9 to 17 for Route 96).

We have also observed certain biases in the cross-route estimation. For instance, there is an underestimation of the average passenger load between stop 14 and stop 20 on Route 11. Stop 20 is in proximity to Northcote High School, which attracts a significant number of school trips. These school trips will not be captured by Route 86 and 96, so the boarding flows for these school trips, which are likely to be generated from a few upstream stops, will be underestimated. This limitation can be overcome by adding additional explanatory variables, such as the presence of a school within walking distance of a stop. However, the limited data sample may not ensure that these types of features appear in both training and test data. For example, the inbound segments of Route 86 and Route 96 do not contain any schools. As a result, the model cannot learn the effect of the schools from the training data and accurately estimate the passenger demand for school trips on Route 11.

Another significant underestimation occurs in the entire inbound segment of Route 96. We speculate that this may be attributed to the lack of features related to mode choice, including the availability and quality of alternative modes. For example, residents living along Route 11 and Route 86 have the option to take the train to the CBD, which is typically faster than trams. Conversely, for stops on Route 96, residents in this area may prefer trams due to the relatively shorter travel time. Consequently, even with an equivalent population, more passengers are observed on Route 96 compared to Route 11 and 86. Therefore, using data from Route 11 and Route 86 to estimate loads on Route 96 will result in an underestimation in the inbound segment.

Although these biases cannot be eliminated by the current models, results may remind transit operators to give specific attention to these trips. These trips are unique to each route and provide valuable information for public transport network planning.

Conclusion and future work

This paper proposes a set of methods to estimate passenger loads in public transport networks where AFC coverage is low and skewed. We propose a set of supervised learning methods based on the existing observations of passenger flows collected by APC systems. Taking into account the coverage of APC data, two estimation problems, utilising the same methods, are considered: (1) same-route estimation for routes with APC systems deployed on a subset of the fleet, and (2) cross-route estimation for routes without any APC systems.

The proposed models and benchmarks are tested on three tram routes in Melbourne, Australia. The results demonstrate that introducing supervised learning models with features developed from exogenous data sources yields superior results compared to estimating passenger flows directly from records in the AFC data. The basic supervised learning model can be further enhanced by modelling relationships between boarding flows, alighting flows, and passenger loads using either the proposed two-stage method or OD-based method. The former calibrates the estimation of alighting flows using the estimated loads at the previous stop, while the latter utilises the alighting probability from AFC data to calculate the alighting flow.

For the cross-route estimation problem, the pure supervised learning model does not perform well since the training and test data come from different routes. The relationship between passenger flows and input features may vary across different routes. This problem is addressed by the proposed segment-based method, which divides the stops of transit routes into inbound, outbound, and peak segment and employs different supervised learning models for each segment. This method is developed based on the theory that only certain aspects of the relationships between passenger flow data and input features are consistent across different routes, and they may vary between traffic analysis zones with different functionalities. This is an important finding in the understanding of the transferability of supervised learning between different transit routes.

The broad implication of this research is that it enables transit operators to continuously monitor service loads for each individual trip using those limited data sources, even in cases where direct observations are unavailable for the entire transit route. The information on passenger loads not only helps operators to accommodate the variability in passenger demand but also assists passengers in travel planning.

Three recommendations for future research are provided. Firstly, the current methods have not incorporated features related to mode choice. The case study reveals that the proposed model underestimates the passenger loads at stops in the inbound segment of Route 96. This underestimation could possibly be attributed to the ignorance of factors that influence passengers' mode choices, such as transit travel time, transport accessibility, and alternative travel modes. Future research could strategically examine the integration of mode choice models into the existing framework.

Secondly, some transit operators have access to AFC data from other public transport modes, including train and bus services. These data sources may capture valuable information during special events. Recent studies have demonstrated different travel modes may exhibit shared demand patterns temporally and spatially within a city (Li et al. 2020; Zhou et al. 2021). However, the study of the inter-effects is still in its early stages.

Thirdly, the proposed segmentation method, which divides the network into inbound, peak, and outbound segments, is only applicable to network consisting of radial lines. For other network configurations, such as circular network, the mechanism of trip generation and attraction on these networks needs to be further studied and strategically integrated into the current segment-based model framework.

Acknowledgements This research is financially supported by the Victoria Department of Transport and Planning (DTP), Cubic Transportation Systems, and iMOVE Australia. The authors would like to express their gratitude to DTP, iMOVE, Cubic, and Yarra Trams for providing the AFC, AVL, APC data, as well as their insights and consultations.

Author Contributions The authors confirm contribution to the paper as follows: Conceptualization: TY, NN Methodology: TY Data curation: TY, Joseph Leong Formal analysis and investigation: TY, NN, ET Writing - original draft preparation: TY Writing - review and editing: TY, NN, ET Funding acquisition: NN, MS Supervision: NN, MS.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Bie, Y., Gong, X., Liu, Z.: Time of day intervals partition for bus schedule using GPS data. Transp. Res. Part C Emerg. Technol. **60**, 443–456 (2015)

Cantillo, A., Raveau, S., Muñoz, J.C.: Fare evasion on public transport: who, when, where and how? Transp. Res. Part A Policy Pract. **156**, 285–295 (2022)

Cats, O.: Dynamic Modelling of Transit Operations and Passenger Decisions. KTH Royal Institute of Technology, Stockholm (2011)

Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

Cheng, Z., Trépanier, M., Sun, L.: Incorporating travel behavior regularity into passenger flow forecasting. Transp. Res. Part C Emerg. Technol. **128**, 103200 (2021)

Chu, K.K.A., Chapleau, R.: Enriching archived smart card transaction data for transit demand modeling. Transp. Res. Record **2063**(1), 63–72 (2008)

Delbosc, A., Currie, G.: Four types of fare evasion: a qualitative study from Melbourne, Australia. Transp. Res. Part F Traffic Psychol. Behav. **43**, 254–264 (2016)

Eady, J., Burtt, D.: Walking and transport in melbourne suburbs (2019)

Gentile, G., Nökel, K.: Modelling public transport passenger flows in the era of intelligent transport systems. Springer Tracts Transp. Traffic **10**, 641 (2016)

Glick, T.B., Figliozzi, M.A.: Measuring the determinants of bus dwell time: new insights and potential biases. Transp. Res. Record **2647**(1), 109–117 (2017)

Gordillo, F.: The Value of Automated Fare Collection Data for Transit Planning: An Example of Rail Transit od Matrix Estimation. Massachusetts Institute of Technology, Cambridge (2006)

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H.M., Attanucci, J.P.: Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. Transp. Res. Record **2343**(1), 17–24 (2013). https://doi.org/10.3141/2343-03

Han, A.F., Wilson, N.H.M.: The allocation of buses in heavily utilized networks with overlapping routes. Transp. Res. Part B Methodol. **16**(3), 221–232 (1982)

Jenelius, E.: Data-driven bus crowding prediction based on real-time passenger counts and vehicle locations. In: 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MTITS2019) (2019)

Kagho, G.O., Balac, M., Axhausen, K.W.: Agent-based models in transport planning: current state, issues, and expectations. Procedia Comput. Sci. **170**, 726–732 (2020)

Khani, A.: Models and Solution Algorithms for Transit and Intermodal Passenger Assignment (Development of Fast-Trips Model). The University of Arizona, Tucson (2013)

Kurauchi, F., Schmöcker, J.-D.: Public Transport Planning with Smart Card Data. CRC Press, Boca Raton (2017)

Li, C., Bai, L., Liu, W., Yao, L., and Waller, S T.: Knowledge adaption for demand prediction based on multi-task memory neural network. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 715–724 (2020)

Li, Y., Cassidy, M.J.: A generalized and efficient algorithm for estimating transit route ods from passenger counts. Transp. Res. Part B Methodol. **41**(1), 114–125 (2007)

Luo, D., Bonnetain, L., Cats, O., van Lint, H.: Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. Transp. Res. Record **2672**(8), 175–186 (2018)
McCord, M.R., Mishalani, R.G., Goel, P., Strohl, B.: Iterative proportional fitting procedure to determine bus route passenger origin-destination flows. Transp. Res. Record **2145**(1), 59–65 (2010)
McNally, M.G.: The Four-Step Model. Emerald Group Publishing Limited, Bingley (2007)
Miller, E., Sánchez-Martínez, G.E., Nassir, N.: Estimation of passengers left behind by trains in high-frequency transit service operating near capacity. Transp. Res. Record **2672**(8), 497–504 (2018)
Moreira-Matias, L., Cats, O.: Toward a demand estimation model based on automated vehicle location. Transp. Res. Record **2544**(1), 141–149 (2016)
Munizaga, M.A., Palma, C.: Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C Emerg. Technol. **24**, 9–18 (2012). https://doi.org/10.1016/j.trc.2012.01.007
Munizaga, M.A., Gschwender, A., Gallegos, N.: Fare evasion correction for smartcard-based origin-destination matrices. Transp. Res. Part A Policy Pract. **141**, 307–322 (2020)
Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M.: Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. Transp. Res. Record **2263**(1), 140–150 (2011)
Nassir, N., Hickman, M., Ma, Z.-L.: Activity detection and transfer identification for public transit fare card data. Transportation **42**(4), 683–705 (2015). https://doi.org/10.1007/s11116-015-9601-6
Neema, N., Hickman, M., Ma, Z.: Statistical inference of transit passenger boarding strategies from farecard data. Transp. Res. Record **2652**(1), 8–18 (2017)
Neema, N., Hickman, M., Ma, Z.-L.: A strategy-based recursive path choice model for public transit smart card data. Transp. Res. Part B Methodol. **126**, 528–548 (2019)
Pelletier, M.-P., Trépanier, M., Morency, C.: Smart card data use in public transit: a literature review. Transp. Res. Part C Emerg. Technolo. **19**(4), 557–568 (2011)
Siebert, M., Ellenberger, D.: Validation of automatic passenger counting: introducing the t-test-induced equivalence test. Transportation **47**(6), 3031–3045 (2020)
Strathman, J.G.: An evaluation of automatic passenger counters: validation, sampling, and statistical inference (1989)
Sun, L., Tirachini, A., Axhausen, K.W., Erath, A., Lee, D.-H.: Models of bus boarding and alighting dynamics. Transp. Res. Part A Policy Pract. **69**, 447–460 (2014)
Sun, W., Schmöcker, J.-D., Fukuda, K.: Estimating the route-level passenger demand profile from bus dwell times. Transp. Res. Part C Emerg. Technol. **130**, 103273 (2021)
Trépanier, M., Tranchant, N., Chapleau, R.: Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transp. Syst. **11**(1), 1–14 (2007)
Turnquist, M.A.: A model for investigating the effects of service frequency and reliability on bus passenger waiting times. Transp. Res. Record **663**, 70–73 (1978)
Zhou, Q., Gu, J., Lu, X., Zhuang, F., Zhao, Y., Wang, Q., Zhang, X.: Modeling heterogeneous relations across multiple modes for potential crowd flow prediction (2021). arXiv preprint arXiv:2101.06954

Mr Tianwei Yin is a PhD student within the Department of Infrastructure Engineering, University of Melbourne. His thesis focuses on using data-driven methods to support public transport planning, with particular interests in demand forecasting and planning methods in public transport.

Dr Neema Nassir is a Senior Lecturer in Transport Engineering and the Discipline Coordinator for Civil Engineering within the Department of Infrastructure Engineering, University of Melbourne. His research is focused on new methods to simulate, model, design and manage public transport, shared mobility, and connected/automated transport systems.

Mr Joseph Leong was a research assistant within the Department of Infrastructure Engineering, University of Melbourne during this project. Skilled in database management, he provided valuable support with data curation.

Prof Egemen Tanin is a professor within the School of Computing and Information Systems, the University of Melbourne. His research interests include spatial databases and mobile data management. His current focus is on traffic data and future management systems in this arena.

Prof Majid Sarvi is the chair and professor in Transport Engineering and the program director of the "Transport Technologies" at the University of Melbourne. His field of research includes Artificial Intelligence in transport, connected and automated transport and Intelligent Transport Systems.