Journal of Public Transportation 25 (2023) 100070

Contents lists available at ScienceDirect

Journal of Public Transportation

journal homepage: www.journals.elsevier.com/journal-of-public-transportation

Estimation of alighting counts for public transportation vehicle occupancy levels using reverse direction boarding

Kamer Özgün [a,*], Melih Günay [b], Doruk Başaran [b], Joseph Ledet [b]

[a] Antalya Bilim University, Turkey
[b] Akdeniz University, Turkey

ARTICLE INFO

Keywords:
Vehicle occupancy
Alighting counts
Offboarding
Destination inference
Smart card

ABSTRACT

In order to improve the performance of a public transport system, it is important to know alighting counts as well as boarding counts at bus stops. While boarding is almost always available through a fare collection system, public transport systems usually do not count alighting passengers. This is either due to the overhead that may be required for passengers at exit or the installation of relatively expensive Automatic Passenger Counters (APC) counters at each vehicle. Therefore, such expensive deployments are mostly not encountered in public-transport systems. In our research, for round trip lines that are balanced in daily passenger counts on both forward and backward routes, the alighting counts of a target route are inferred using only the daily boarding data. Vehicle occupancy levels are determined on a trip basis owing to the characteristic boarding pattern of each line. The validity of the proposed method was determined and verified using video recordings of arbitrarily selected trips. Consequently, it may be possible to modify scheduling algorithms to improve vehicle fleet utilization and increase passenger comfort in public transportation.

1. Introduction

Today, Automatic Fare Collection (AFC), Automatic Passenger Counter (APC), Automatic Vehicle Positioning (AVL), and Global Positioning Systems (GPS) are powerful data sources available to researchers. Public transit data mining has provided significant opportunities and accelerated research potential to improve operational efficiency together with service quality (Harrison et al., 2020; Lu et al., 2020; Welch and Widita, 2019; Zhu et al., 2019; Li et al., 2018; Pelletier et al., 2011).

The most commonly studied attributes related to the quality of public transportation are (Redman et al., 2013); reliability (i.e. the matches between actual service and route timetable), frequency (related to waiting times at bus stops), travel times (related to time spent in the vehicle), accessibility and network connectivity (Bulut et al. (2021)). The goal is to operate a limited number of vehicles on the right routes at the right times while keeping the transportation system within the optimized boundaries of the performance measures (Stewart et al. (2016)), while meeting the demand that is constantly changing during the day and even on a per route basis (Özgün et al. (2021a)). Among several other metrics presented recently by Özgün et al. (2021b), vehicle occupancy levels are directly related to both service efficiency and service quality.

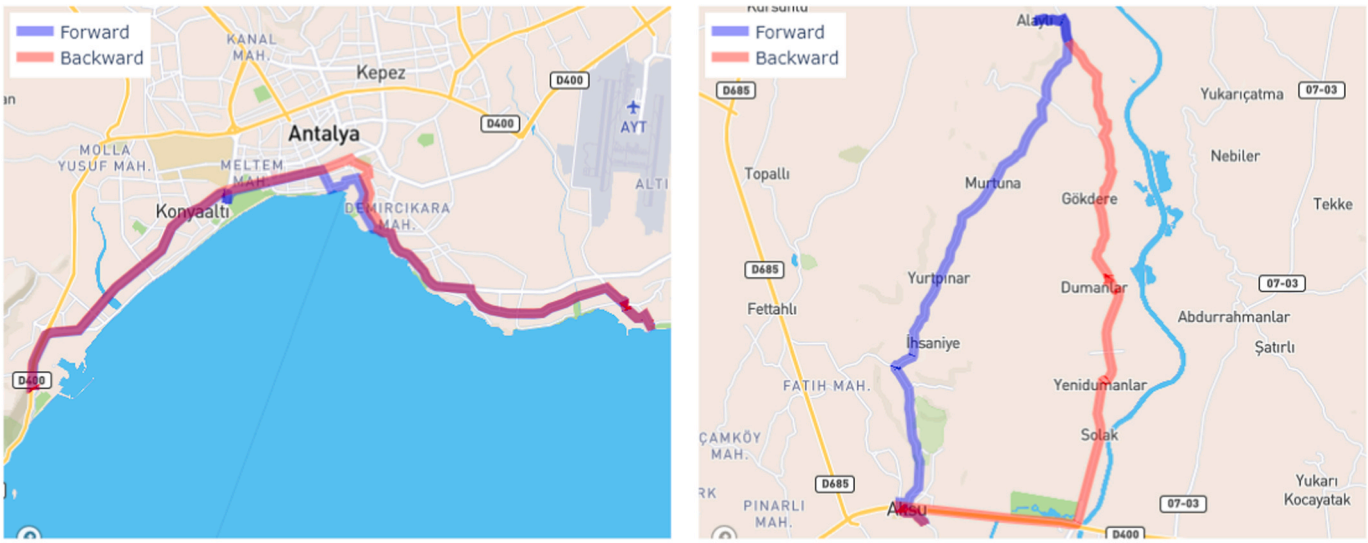
Crowded vehicles may demonstrate faulty frequency settings (i.e. the number of trips per time needed to satisfy the passenger demand) and poor service quality, while empty vehicles may signify operational inefficiency. Determination of vehicle occupancy levels are crucial for the control and the management of limited vehicles in public transport operations (Bertsimas et al., 2020; Lee et al., 2021).

Accurate occupancy estimation enables transportation authorities to optimally allocate the appropriate number and type of vehicles for a given route at calculated frequencies. Such allocation of resources enables vehicles to be deployed where they are most needed. This, in turn, may result in the reduction of the overall number of vehicles on the road resulting in reduced carbon emissions and contributing to sustainability. Furthermore, vehicles are less likely to be overcrowded improving passenger safety and comfort.

Occupancy levels can be determined by tracking on- and off-boarding at stops or stations (Yang et al. (2021)). Although on-boarding data are usually available through an AFC system, alighting is rarely counted (Ma (2013)). APC systems may estimate the occupancy levels of vehicles by deploying somewhat expensive technologies such as

* Corresponding author.
E-mail address: kamer.ozgun@antalya.edu.tr (K. Özgün).

https://doi.org/10.1016/j.jpubtr.2023.100070
Received 23 July 2022; Received in revised form 10 October 2023; Accepted 12 October 2023
Available online 21 October 2023



(a) Point-to-Point Network Example (*KL08*) (b) Ring Network Example (*AC05*)

Fig. 1. Common Public Transportation Topologies for Antalya.

stereo vision and vehicle dynamic weight sensors, or relatively inexpensive technologies such as infrared beam sensors and GPS data (Lu et al., 2020; Nassir et al., 2011; Mohammed and Oke, 2022). However, it is important to note that technology adoption varies across regions and transportation networks. Some public transportation systems have embraced APC systems more extensively than others have. Nevertheless, even today, most public transportation in underdeveloped and developing countries does not have APC systems installed GlobeNewswire (2022).

Furthermore, some mobile applications track both boarding and alighting passenger counts via crowd-sourcing (Glasgow, 2023; Steinfeld, 2023; Gannes, 2012). Tracking bus fullness through mobile phones is one of the most critical metrics used to quantify public-transportation service quality and consumer satisfaction. Often, it is simply reported by passengers in real-time. However, human perception does not always reflect a precise number of passengers (Pi et al. (2018)).

For strategic and tactical transit planning, real-time data are often unnecessary (Pelletier et al. (2011)) and the literature on the estimation of unknowns in public transport is still emerging (Li et al., 2018; Hussain et al., 2021)). Faroqi et al. (2018) summarizes studies published after 2010 on smart card data and group them based on their applications as mining travel patterns and estimating trip purpose. Although real-time crowdness information is useful, it is still possible to enrich real-time crowdness information with better alighting prediction and estimate future bus fullness on the existing route.

In the current literature, focusing on entry-only AFC systems, alighting location estimation, also known as the destination inference problem, can be classified into trip-chaining (linked), probabilistic (mostly unlinked), and Artificial Neural Network (ANN) models (Li et al. (2018)). Li et al. (2018) discussed the advantages and disadvantages of destination estimation models. According to Li et al. (2018), ANN-based destination estimation models require large amounts of data and numerous variables to be captured for accurate training, which makes these systems complex. Furthermore, these models can only be applied to systems in which entry-exit information is collected through the APC. On the other hand, probability models may infer total on-off passenger counts with limited information while providing anonymity in the estimations.

In the trip-chaining model, alighting locations are inferred from a particular passenger's successive trips (Ma et al. (2013)). The deficiencies of the trip-chaining model may be stated as follows:

- The destinations of transactions with exactly one trip per day cannot be inferred using one-day smart card data (Hussain et al. (2021)).
- Destination inferences are not over aggregate boarding data, but over a small dataset obtained by monitoring individuals with two or more trips.

The main idea of probabilistic models is to calculate the alighting probability of passengers on possible destination stops by considering the travel distance and passenger counts on other stops and storing the results in an origin-destination (OD) matrix. Unlinked probabilistic models not inferred from individuals (Ait-Ali and Eliasson (2019)) define system entropy and maximize it under predetermined conditions. In Yang et al. (2020) a nonlinear programming model within a deep optimization framework was used with 5-day of boarding data by using uniform 10 min time intervals without considering the opposite directions. In Cheng et al. (2021), individual boarding data were characterized by a multinomial distribution using real or pre-estimated alighting data for training. In Cheng et al. (2021), the deployed probabilistic model was derived from Latent Dirichlet Allocation, which is generally used in Natural Language Processing. The probabilistic model was used as a continuation of a rule-based model based on the trip-chaining algorithm. Their rule-based model consisted of three rules: the first was the standard tripchaining procedure, the second was the prediction of the last destination from the first origin of the same day, and the third was the prediction of the last destination from the first origin of the following day. The destinations of linked passenger trips, which were 85.26% of their case study, were estimated using this rule-based model. In short, Cheng et al. (2021) estimated the destinations of unlinked passenger trips using estimations of linked passenger trips along the same route. However, their model is complex owing to the rules employed.

This study involves the development of a novel probabilistic approach that incorporates reverse route boarding by utilizing a one-day smart card dataset for a specific service line. To the best of our knowledge, this is the first study to suggest inference estimations based on reverse direction. The results indicate that the daily dataset consisting of linked passenger trips can be used to determine potential destinations at the route level.

To enhance anonymity while decreasing complexity, time-series clusters of single-day smart-card data are aggregated for a particular route (Özgün et al. (2021a)), and the most suitable time slot of the

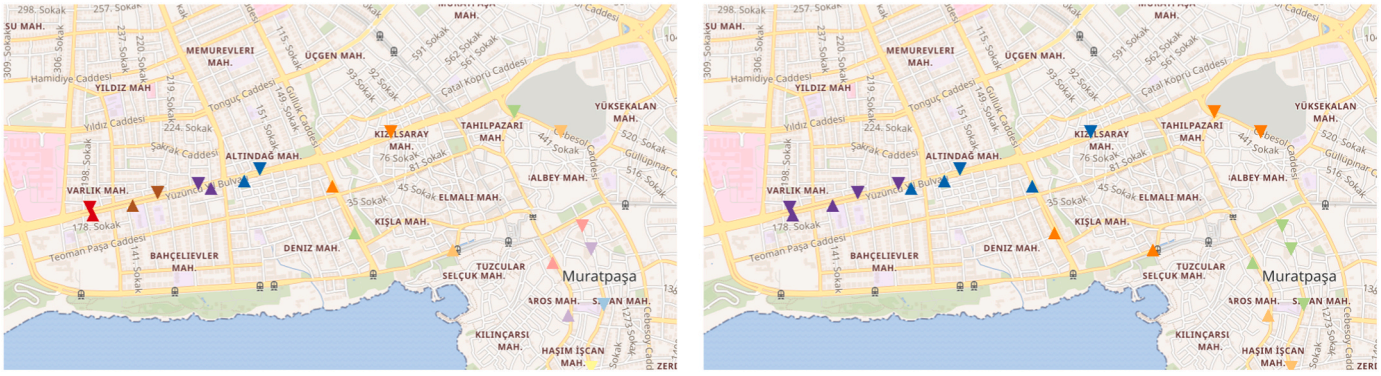


Fig. 2. Comparison of Direct VBSPs and Aggregated VBSPs on a Segment of Bus Line *KL08*.

reverse route is used as an unlinked dataset to estimate alighting counts.

The proposed linked and unlinked approaches were compared with the actual alighting counts obtained from video recordings of arbitrarily selected bus trips instead of relying on costly manual survey data (Farzin (2008)). The accuracy of the estimations was evaluated using several error measures, as detailed in Results and Discussion section.

The outcomes of this research provide valuable insights for transit agencies and policymakers, enabling them to determine the load of each route throughout its journey and make more informed decisions regarding the provision of public bus transit services.

2. Methodology

This study focused on bus transportation, which is preferred by large cities. However, the proposed alighting count estimation methods that use boarding data in the reverse direction may also be applied to other modes of transportation. The primary data source is the Department of Transportation in the City of Antalya, Turkey. This study is based on smart card data collected on a typical weekday during the pre-Covid-19 period. The dataset consists of Passenger ID, Bus Stop ID (passenger's boarding stop), Boarding Time, Bus ID, Line ID and Route ID over 305 bus lines and 608 routes with a total of 7347 trips (single direction services) carried a total of 381,962 passengers. Most of the analyses were performed using Knime software (Berthold et al. (2007)) and custom-developed Python scripts.

Destination inference may be challenging when multiple network typologies (a: star, b: Hub and Spoke, c: Point-to-Point Direct, d: ring) are considered (Taaffe (1996)). In Antalya, the public transportation network topology may be considered mostly Point-to-Point and Hub & Spoke (See Fig. 1). Only six out of 115 lines (≤%5) have the ring topology, which also primarily includes sparse routes. Circular bus routes are usually around a central area or neighborhood, and are suitable for providing transverse links between suburbs or satellites, avoiding congested centers, and filling the service gap left by existing public transit (Taaffe (1996)). Furthermore, sparse routes do not often suffer from overcrowding. Therefore, this study focuses on the backbone lines in public transportation.

To estimate the alighting counts using the boarding data in the reverse direction, the following trip features are required: .

- A point-to-point network topology with parallel lines running against each other in both directions is shown in Fig. 1a. Note that the Ring Topology presented in Fig. 1b is unsuitable for this study.
- The dayend passenger counts (Özgün et al. (2021b)) in both directions are balanced.

A bus route is a sequential list of stops in either the forward or backward direction. Unlike subway lines, reverse bus routes on urban highways may have different route lengths, and bus stops are not necessarily lined up reciprocally. Considering a single bus line, suppose that N_F and N_B are the number of bus stops in the forward and backward directions, respectively. The GPS coordinates of bus stops help determine their physical proximity. Then, virtual bus stop pairs, VBSP with an index same in both directions are created mutually. Creating VBSP also eliminates the need for high stop symmetry.

This paper describes two different approaches to creating VBSPs for a particular bus line. The algorithms are explained in Sections 3.1 and 3.2 and the results of the two approaches are shown in Fig. 2 as an example of an arbitrarily selected busy bus line *KL08*. Note that in both approaches, the number of VBSPs is equal to $N \leq min\{N_F, N_B\}$, and each VBSP has its own index number; thus, one of the routes has a reversed sequence of VBSP indices, as shown in Fig. 2a.

In this study, the boarding data of the backward direction is used to estimate the alighting counts of the forward line. We applied 2 approaches; i) simply using the daily passenger counts on each stop, ii) custom identified peaks determined by clustering the boarding patterns of a bus line fine-tuned for temporal and spatial corrections. We show that using traditional rush hour peaks of a line may not produce accurate results due to characteristics of a bus line discussed in Özgün et al. (2023).

2.1. Origin-destination probability matrix

The first task of the proposed approach is to calculate the alighting probabilities from each bus stop to the subsequent bus stop along the path. The probability that a passenger boarded at bus stop $i \in \{1, 2, \ldots, (N-1)\}$, alights at a bus stop $j \in \{(i+1), (i+2), \ldots, N\}$ along the direction F is denoted by $P_F(i, j)$. Let $C_B(j)$ denote the boarding counts at bus stop j in the backward direction B. The probability $P_F(i, j)$ is then calculated using Eq. (1) as follows:

$$P_F(i,j) = \frac{C_B(j)}{\sum_{j=i+1}^{N} C_B(j)} \tag{1}$$

The next task is to store the probabilities from the origin bus stop to candidate destination bus stops on the path in a matrix called the Origin-Destination Probability Matrix. The lower triangular matrix provides the alighting probabilities, P_F where the alighting stops are in columns and the boarding stops are in rows. Note that the probabilities in each row of P_F add up to 1.

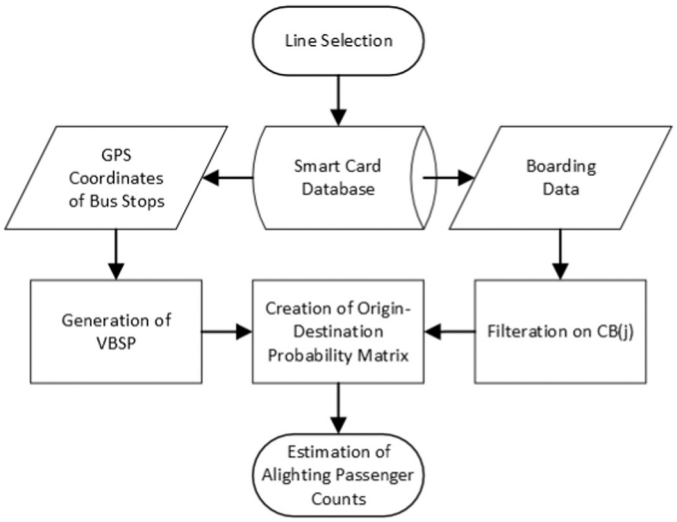


Fig. 3. Overall Alighting Passenger Counts Estimation Workflow.

$$P_F = \begin{matrix} 0 & P_F(1,2) & P_F(1,3) & P_F(1,4) & P_F(1,5) & \dots & P_F(1,N) \\ 0 & 0 & P_F(2,3) & P_F(2,4) & P_F(2,5) & \dots & P_F(2,N) \\ 0 & 0 & 0 & P_F(3,4) & P_F(3,5) & \dots & P_F(3,N) \\ 0 & 0 & 0 & 0 & P_F(4,5) & \dots & P_F(4,N) \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \dots & \dots & P_F(N-1,N) \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 \end{matrix} \quad (2)$$

2.2. Estimation of alighting passenger counts

Let $E_F(j)$ denote the estimation of alighting counts along the forward direction at bus stop j. Once the alighting probability matrix is constructed, the total boarding count is multiplied by the alighting probability of each bus stop along the forward path, as given by Eq. (3):

$$E_F(j) = \sum_{i=1}^{j-1} C_F(i) P_F(i,j) \quad (3)$$

By substituting Eq. (1) into Eq. (3), we have obtained:

$$E_F(j) = \sum_{i=1}^{j-1} C_F(i) \frac{C_B(j)}{\sum_{j=i+1}^{N} C_B(j)} \quad (4)$$

2.3. Calculation of vehicle occupancy levels

The vehicle occupancy level at a particular bus stop or VBSP area $s \in \{1, 2, \dots, N\}$ in direction F, denoted by $O_F(s)$, is calculated as the difference between the number of boarding and alighting trips. Eqs. (3) is used to calculate the alighting estimates for each VBSP. Let $A_F(j)$ denote the actual alighting counts in a particular VBSP area $j \in \{1, 2, \dots, N\}$ in the direction F. Then,

$$O_F(s) = \sum_{j=1}^{s} C_F(j) - A_F(j) \quad (5)$$

3. Alighting estimates approaches at route level

Destination inference using smart card data is an important and widely studied problem. Existing methods primarily exploit the continuity of trips (linked methods) and infer destinations based on assumptions or rules, particularly to identify relevant transfer connections from entry-only AFC records at the network level.

In this study, we demonstrate that a one-day smart card dataset filtered by individual passenger records results with a linked dataset can be used to determine potential destinations at the route level. Linked methods are explained in details in sections 3.1 and 3.3.

On the other hand, considering a single line, the time lag between getting on and off a particular stop may vary significantly among passengers based on various trip purposes. In Özgün et al. (2023), it was shown that the number of peaks, the start and end times of peaks, and the duration of peaks change at the route level (from bus line to bus line) in the transit network. Peak hours identified by clustering boarding patterns provides temporally and spatially tuned, aggregated,thus unlinked dataset. Unlinked M2 method is presented in section 3.2.

The workflow of the proposed approach is illustrated in Fig. 3. It should be noted from Fig. 3 that the estimation methods differ from each other in terms of the value of N obtained by different VBSP generation methods and time slot or round-trip filters on $C_B(j)$ used in Eq. (4). The three approaches considered for the analysis are described in detail in the following sections.

3.1. M1: Direct VBSPs with round trip passengers (Home-based)

Commuters primarily use public transportation for i) going to work regularly, ii) visiting hospitals or shopping malls occasionally, iii) leisure and sightseeing, or iv) shopping trips for essential needs. Among these reasons, the largest contributor to public transit system volume is mainly due to commuters who work regularly.

The aim is to show that the daily data set consisting of round trip passengers can be used to determine potential destinations at the route level. The assumption is that the passenger boards the bus and disembarks for the business at a stop which they then use to embark on the return after some time. Under that assumption, round-trip bus commuters taking only 2 trips on both directions of the same line are filtered out in method M1. Although there are not many of these types of passengers, it is still possible to obtain sufficient statistics that help to estimate alighting counts.

Method M1 runs over the direct VBSPs with round-trip passengers. Thus, M1 is a home-based model uses individual passenger counts i.e. it is a linked method. The method M1 reverses the route directly by arranging opposing bus stops by their sequence number. Let ordered pair (f,b) denotes pairing of f^{th} bus stop on direction F with b^{th} bus stop on direction B. If $|N_F - N_B| = 0$ then $N = N_F = N_B$ and then ordered pairs are simply (1, N), (2, N − 1), …, (N, 1). If $|N_F - N_B| > 0$, then VBSP count becomes equal to $N = max\{N_F, N_B\}$ and pairing is slightly dislocated by ΔI in Eq. (6), by shifting bus stops of N_B to the center. In that case, the last bus stop of the shorter route will be paired with the $(\Delta I+1)^{th}$ bus stop of the longer route.

$$\Delta I = \left\lceil \frac{|N_F - N_B|}{2} \right\rceil \quad (6)$$

Without loss of generality, suppose that $N_F > N_B$ then first pairing is $((\Delta I + 1), N_B)$ and last paring is $((N_F - \Delta I + 1), 1)$. Dummy bus stops are assigned zero boarding counts i.e. $C_B(j) = 0$ for $j \in \{1, 2, \dots, \Delta I\} \cup \{(N_F - \Delta I + 1), \dots, N_F\}$.

3.2. M2: aggregated VBSPs with inversed time slot (modified home-based with time lag)

A passenger alights at the stop closest to the destination within walking distance W. This also applies to the boarding passes. Under this assumption, instead of pairing the opposing stops, the bus line is divided into zones by a predetermined W, and the bus stops are aggregated into groups in these zones to obtain the opposing VBSPs.

Method M2 runs over aggregated VBSPs with anonymous passenger counts and is an unlinked method. This is a modified home-based model that uses passenger counts with a particular time lag determined using a time-series clustering algorithm.

Let, L_F and L_B denote route lengths in forward and backward

directions respectively.

$$N = \left\lceil \frac{\min\{L_F, L_B\}}{W} \right\rceil \tag{7}$$

The number of VBSP on a particular line is the ceiling for the division of its shorter route by a given walking distance. In Eq. (7), the parameter W may be modified in order to adjust the reciprocity of VBSPs. Note that the zone length (i.e., the distance between the first and last stops in a particular VBSP) is always less than or equal to W, which is set to 1 km in our study. Let the zone lengths F and B be denoted by Z_F and Z_F respectively.

Once the number of VBSP (N) has been determined by Eq. (7), bus stops are consolidated into groups based on their VBSP index. Each VBSP has a corresponding reverse bus stop(s) on the reverse route with the same index.

VBSP range for each route are calculated. The bus stops' intervals along the route are precisely measured from the starting point of each route. In accordance with the specified regulations, each Bus Stop is designated to a specific VBSP, as outlined in Algorithm 1.

Algorithm 1. Bus Stop Aggregation.

Algorithm 1 Bus Stop Aggregation

```
procedure BSA(LineID)
    N ← ⌈min{L_F, L_B}/W⌉                                   ▷ Eq. 7
    Zones ← Array(size:2)
    Zones[1] ← Array(size:N_F)
    Zones[2] ← Array(size:N_B)
    for each r ∈ Direction do
        Z_r ← L_r/N                                          ▷ Z_F and Z_B
        i ← 1
        while i ≤ N_r do                                     ▷ Each Bus Stop of Route
            Z_ri ← LocationOnRoute(r, i)                     ▷ Distance from Starting Point of Route
            if r is F then
                Zones[1][i] ← ⌊Z_Fi / Z_F⌋
            else if r is B then
                Zones[2][i] ← N − ⌊Z_Bi / Z_B⌋ + 1
            i ← i + 1
    return Zones
```

In general, the peak activity hours occur in the morning and afternoon. The peak is caused by regular employees traveling from residential centers located on the outskirts to business centers in a city during the morning rush hour, and vice versa during the late afternoon rush hour. As a result, it is likely that a public transport passenger will take the reverse route of the same line to return to their origin with a time lag Özgün et al. (2023). The assumption is that the load of nonworking commuters may be distributed randomly throughout the day. Under these assumptions, morning alighting counts can be estimated from boarding counts during the afternoon rush hour, and vice versa.

In a typical public transport network, the occurrence of peak and off-peak hours is not uniform for all paths Özgün et al. (2021a). The constantly changing passenger activities make it challenging to establish fixed and common rush hours in the morning and afternoon. To address this, a stepped time slot clustering (STSC) algorithm is employed in this study to identify the peak hours of the morning (AM) and afternoon (PM) as reverse time slots based on passenger boarding data. The STSC algorithm is explained in detail in Özgün et al. (2021a).

In method M2, for example, boarding counts in the set $C_B(j)$ are filtered under consideration of a time effect to determine $C_{B(PM)}(j)$ counts of boarding for VBSP j on direction B in afternoon (PM). For M2, previous equations (1), (4) are modified as follows:

$$P_{F(AM)}(i,j) = \frac{C_{B(PM)}(j)}{\sum_{j=i+1}^{N} C_{B(PM)}(j)} \tag{8}$$

$$E_{F(AM)}(j) = \sum_{i=1}^{j-1} C_{F(AM)}(i) \frac{C_{B(PM)}(j)}{\sum_{j=i+1}^{N} C_{B(PM)}(j)} \tag{9}$$

3.3. MTC: modified trip chain

In this investigation, the trip chaining method was adapted to process a one-day smart card data set and incorporate reverse route boarding. As the number of round-trip passengers is a subset of the trip chaining method, the MTC approach estimates over a more extensive dataset, including instances with multiple boarding passengers, in comparison to the M1 method. Assuming that each instance of boarding data consists of:

- *Passenger Id* or *Smart Card Id*
- *Route Id* as a combination of *Line Id* and *Direction*
- *Boarding Bus Stop Id*
- *Boarding Time*

A passenger who has done T trips in a day has a *Boarding Bus Stop Id* array denoted by *trips*[T]. Let $t \in \{1, 2, \ldots, T\}$ denotes the indices for *trips*

[T]. For example, when $t = 2$, *trips*[2] points to the second boarded bus stop of a particular passenger.

In the trip chaining method, the alighting bus stop of a trip t is estimated via the location of the boarding bus stop of the next trip $t + 1$. Without loss of generality, consider again the direction F of a particular line. If $\exists$ t points to $i \in \{1, 2, \ldots, (N-1)\}$, the alighting bus stop $j \in \{(i+1), (i+2), \ldots, N\}$ could be the closest bus stop to $t + 1$. Here, the distance between $t + 1$ and j has to be lower than the predetermined walking distance W, otherwise *trips*[T] of that passenger is ignored. In order to estimate the alighting bus stop of the last trip of a passenger $t = T$, the boarding bus stop of the first trip of that passenger $t = 1$ is used under the assumption that the passenger returns back to their origin when using the daily smart card data.

Estimations from each passenger's successive trips result with boarding-alighting pairs (i, j) that include information about passenger id, route id, boarding and alighting bus stops. Those pairs are filtered by the target route appropriately into the OD matrix by using either direct or aggregated VBSPs.

Let $C_T(i, j)$ denote counts of passengers which board at i and estimated to alight at j. Then the probability matrix is obtained by using Eq. (10) for each cell where $j > i$.

$$P_F(i,j) = \frac{C_T(i,j)}{\sum_{j=i+1}^{N} C_T(i,j)} \tag{10}$$

$$E_{F(T)}(j) = \sum_{i=1}^{j-1} C_F(i) \frac{C_T(i,j)}{\sum_{j=i+1}^{N} C_T(i,j)} \tag{11}$$

4. Results and discussion

This paper proposes a novel probabilistic approach to estimate the alighting counts for entry-only AFC systems on a route-by-route basis. A recent study conducted by Cheng et al. (2021) claims that based on the origin station and departure time, the destination can be determined through statistical analysis. In their model, the departure time is discretized into 1-hour intervals, which is a common practice adopted in the literature. Actual or pre-estimated alighting counts are required to train their inference model. In contrast, our proposed method does account for the dynamic, route-specific nature of passenger demand (Özgün et al., 2023) by clustering time slots for each route, as explained in Özgün et al. (2021a). The proposed method does not rely on actual alighting counts; instead, it generates alighting estimations for each route using only actual boarding data on opposing routes that are already available in a typical smart card dataset.

Compared with the ANN-based approach, a probabilistic approach provides a much simpler alternative that requires fewer data and variables. For instance, Jung and Sohn (2017) suggests an ANN approach that necessitates 27 variables. Although certain parameters can be deduced from the boarding data, land use and socioeconomic data must be obtained from external sources. The techniques outlined in this paper differ in terms of the data sources required, as they are based only on filtered boarding counts assigned to Virtual Bus Stop Pairs (VBSPs) derived from GPS coordinates of bus stops (See Fig. 3).

The evaluation process of the estimation methods is shown in Fig. 4. Several measures can be used to compare the performances of the proposed approach. The measures utilized in this study are based on the deviation between the "actual" values and "the value estimated by one of the proposed methods" for each VBSP area over a route. These are the mean absolute error (MAE) and root-mean-square error (RMSE). We prefer entirely difference-based and not relative measures (such as the symmetric mean absolute percentage error or the log of the accuracy ratio), because we have both the actual and estimated values at zero. Instead, the accuracy percentages based on the overall counts are presented.

The results and discussions are based on the bus line *KL08* of Antalya.

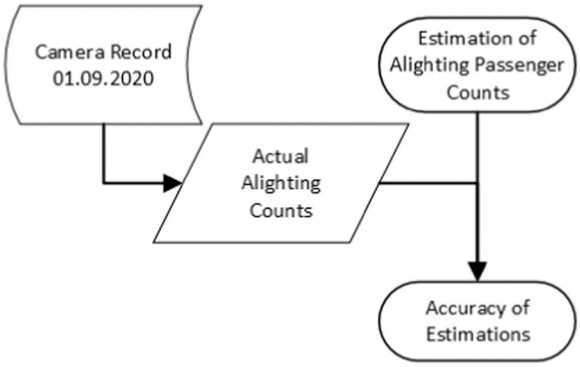


Fig. 4. Evaluation process of estimation methods.

Real life data consists of video recordings of the two most crowded *KL08* forward route trips that took place on September 1, 2020. Passenger activities at each bus stop were counted visually from the video recordings, and the actual bus occupancy counts were obtained. It should be noted that the video recording dates differ from the smart card data dates used for alighting estimations. In this way, the performance of the estimation methods can be measured by comparing their outputs and actual counts.

The results indicate that the proposed approach provides fairly accurate estimates using only the boarding data for entry-only AFC systems, and can be swiftly implemented because of its simplicity and the absence of complex rules.

4.1. Effect of time on estimations

Using day-end passenger counts of *KL08*, on December 18, 2019, as a typical weekday, we have a data set consisting of 9724 passengers in the backward direction and 9745 passengers in the forward direction. Six time slots were identified for *KL08*, as explained in Section 3.2. Because of the low passenger counts early in the morning and late in the evening, the first two and last two slots are merged. Consequently, the four time-cluster boundaries, *Morning*, *Noon*, *Afternoon*, and *Evening* are obtained. Note that for another bus line, the number and length of these slots may change Özgün et al. (2021a).

For *KL08*, the bus stop locations in the reverse direction nearly overlapped. Using Procedure 1, 26 VBSPs are obtained. Once the time clusters and VBSPs are established, the alighting probability matrix (Eq. (2)) is created, commuters of reverse direction are distributed to the corresponding VBSP.

The passenger counts for each bus stop at each direction is plotted and overlaid for various combinations of the selected lines as shown in Figs. 5 and 6 to represent direction effect without and with time effect respectively on boarding counts. By considering the symmetry in Figs. 5 and 6, the assumption that alighting inferences can be based on boarding in the opposite direction is strengthened.

In order to estimate MorningForward alighting counts, the method M2 (Section 3.2) is applied with MorningBackward, NoonBackward, AfternoonBackward, and EveningBackward boarding counts, so that the effect of the time slot on the accuracy of estimations is investigated and summarized in Table 1. The table presents the start and end times of each time slot, the data size on the reversed route (as the total number of passenger counts), and RMSE over 26 VBSPs (as the average deviation in terms of the number of passengers. Considering RMSE column and Afternoon row of the table, it can be concluded that, the deviation from the actual values is 4.34 passengers on average for each VBSP (approximately each kilometer) of the bus line. Thus, with the lowest deviation, time slot *Afternoon* is the most appropriate for the instance on hand.

Consequently, using the traditional uniform rush hour peaks of a line may not produce accurate results. Instead of assuming a fixed rush hour or a single afternoon peak in the estimations, the appropriate time slot

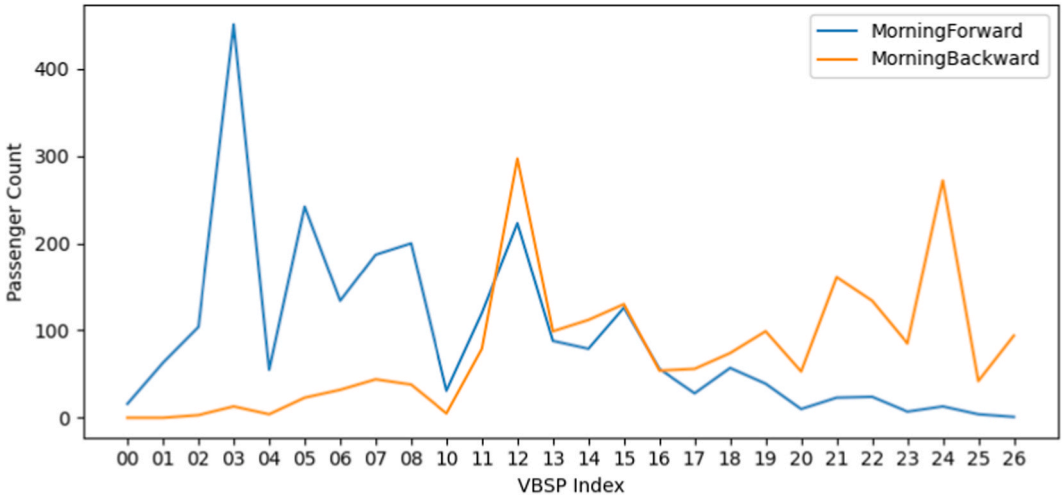


Fig. 5. Morning Forward and Morning Backward Passenger Boarding Counts.

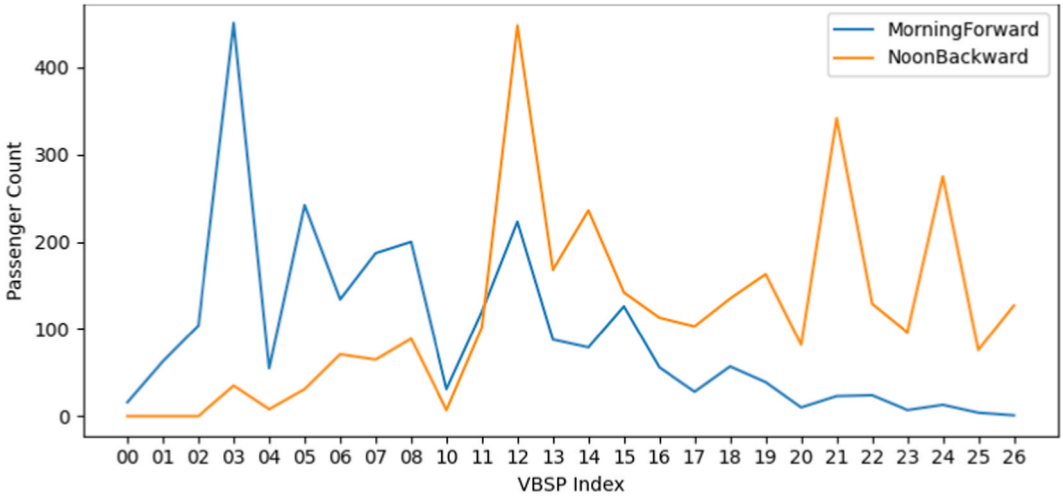


Fig. 6. Morning Forward and Noon Backward Passenger Boarding Counts.

Table 1
Effect of Time Slots in The Method M2.

Time Slots	***Start Time***	***End Time***	***Data Size*** $\sum_{j=1}^{N} C_{B(PM)}(j)$	***RMSE***
Morning	Start of Day	11:00 am	2003	7.00
Noon	11:00 am	3:30 pm	3043	5.76
Afternoon	3:30 pm	8:00 pm	3578	4.34
Evening	8:00 pm	End of Day	1101	7.18

can be determined using clustering algorithms.

4.2. Comparison of estimation methods

In Fig. 7, for *KL08*, the actual alighting values obtained from the video records are indicated by the crosses.

All estimation methods follow similar estimation patterns, but M2 generally estimates better than the other methods. Note that all the methods failed to closely estimate the last VBSP of *KL08*, which can be accepted as an outlier.

This observation is verified by considering a box and whisker plot of the errors, as shown in Fig. 8 which represents alighting estimation errors relative to actual alighting counts in terms of the number of passengers for each method, M1,M2, and MTC. According to the plot, M2 has the median closely aligned with zero error and the lowest variance; thus, it has the most accurate predictions.

The deviations from the actual values as the number of passenger counts are summarized in the first and second rows of Table 2 as the average deviations for all VBSPs. According to the table, M2 has the most accurate results for the alighting count estimation because it resulted in the smallest RMSE and MAE. Accuracy percentages are calculated by considering the ratio of the correct estimates. For example, out of the 224 alightings, 182 are correctly estimated by M2 result with 81% accuracy. Notably, M2 has the highest accuracy percentage.

4.3. Estimation of Vehicle Occupancy Levels

The bus occupancy levels are estimated as explained in Section 2.3 and are presented in Fig. 9. The black line in Fig. 9 indicates the actual occupancy levels. For initial VBSP zones, M1′s estimates are the best, whereas M2 performs better in later VBSP zones.

Finally, the alighting inferences obtained using the proposed methods are discussed for the most populated and balanced bus lines in Antalya, which are determined based on the following selection criteria:
.

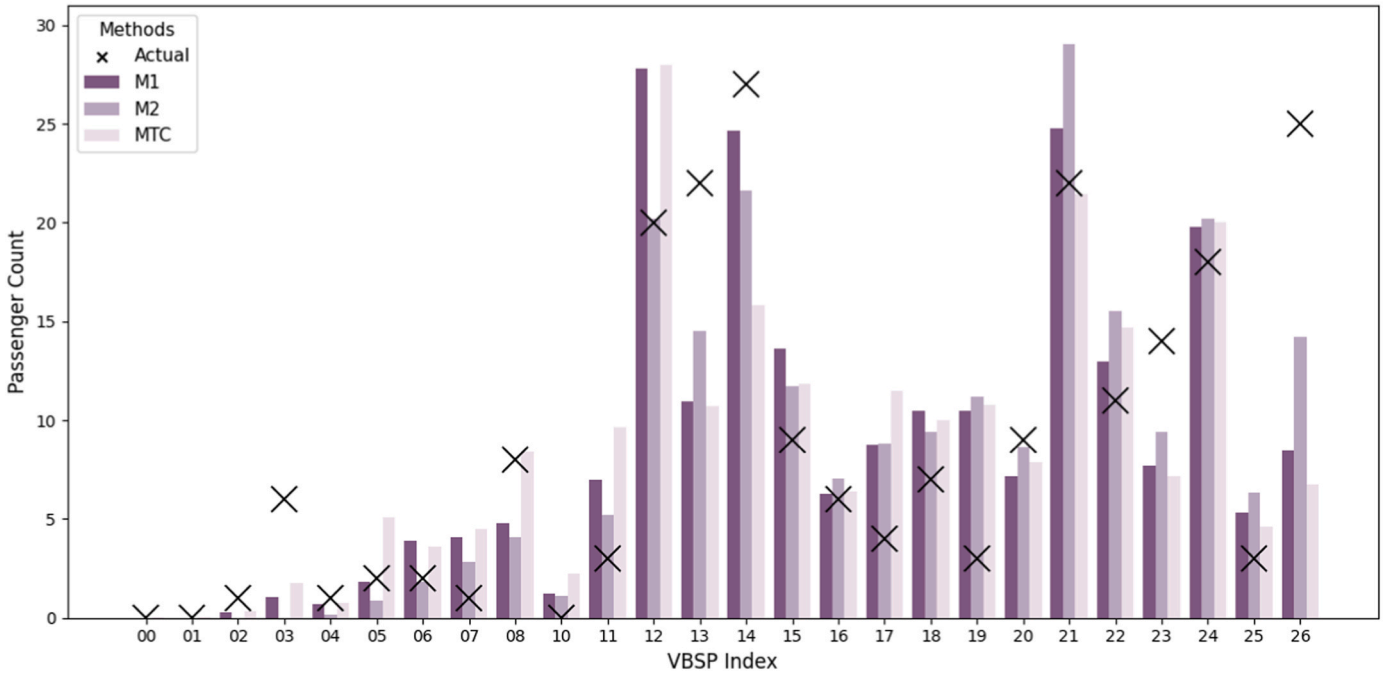


Fig. 7. Alighting Estimation for *KL08* with Actual Counts.

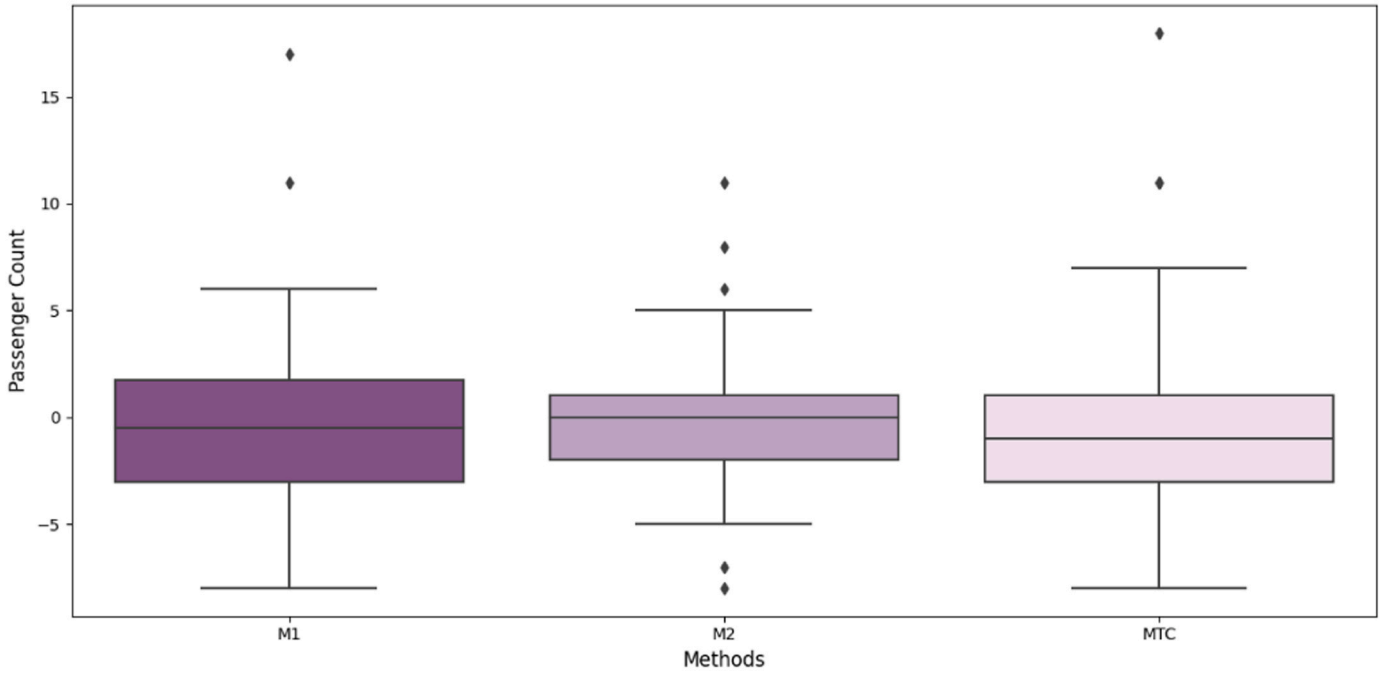


Fig. 8. Alighting Error Comparison for *KL08* on Actual Passenger Counts.

Table 2
Analysis of Alighting Estimation Errors.

	Methods		
Statistics	***M1***	***M2***	***MTC***
RMSE	5.23	4.34	6.04
MAE	3.62	3.19	4.23
Accuracy	79%	81%	76%

- Bus lines having more than 3000 boarding counts on both directions for all day.
- The boarding count difference between reverse directions should not exceed % 10 of maximum boarding count of the either direction.

The comparisons are on the forward directions of the arbitrarily selected bus lines *KL08, KC06, LC07, LF10, UC11* and *VL13* that meet these criteria using a one-day smart card dataset without any travel restrictions or lock-downs. Note that *KL08*, the only line with video recording data, is the basis for our comparisons.

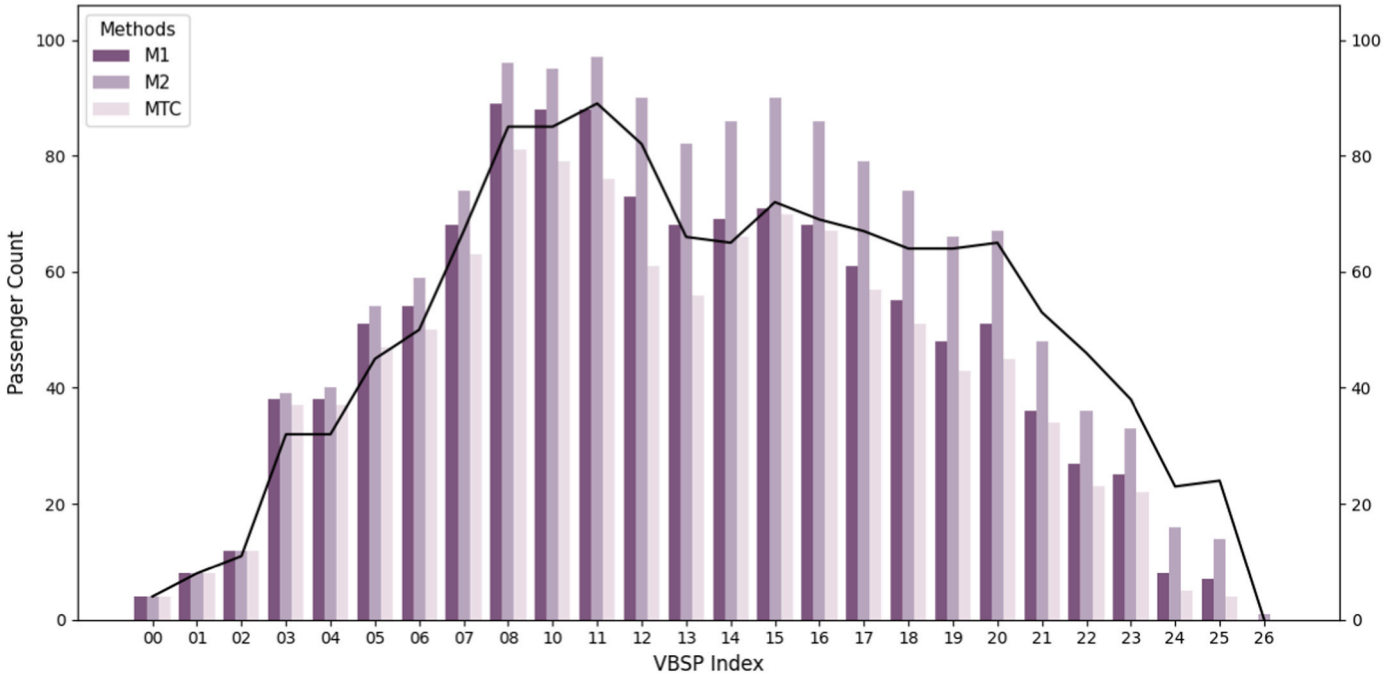


Fig. 9. Bus Occupancy for *KL08* with Actual Counts.

Figure 10 represents estimations of vehicle occupancy levels based on alighting estimation counts for various bus lines. When actual alighting counts are unavailable, the value of $A_F(j)$ in Eq. (5) may be substituted with the expected alighting counts obtained using Equations (4), (9), or (11), depending on the method employed (M1, M2, or MTC).

The black plot in Fig. 10 depicts the estimation when the value of

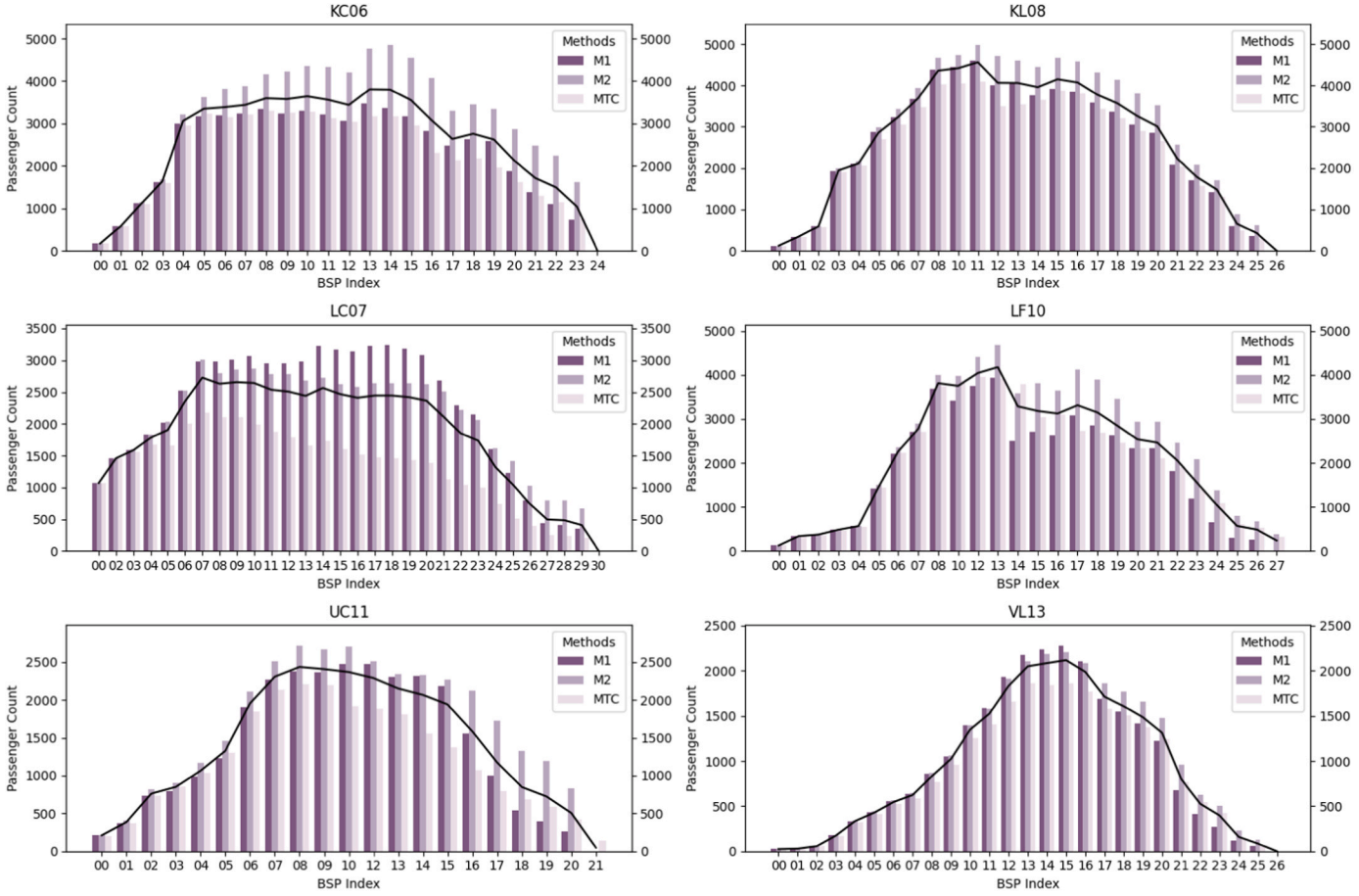


Fig. 10. Occupancy Estimations of All Selected Lines.

$A_F(j)$ is substituted by the three-method average in Eq. (5). On bus line KC06, the occupancy estimates derived from method M2 are notably higher, whereas estimates obtained through methods M1 and MTC are comparable to each other. On bus lines LC07 and UC11, all three methods produce distinct results, with MTC generating significantly lower occupancy estimates. In general, the occupancy estimates obtained through method M2 tend to be higher, while those obtained through method MTC tend to be lower. However, all three methods follow similar trends, as evidenced by the mean line.

5. Conclusion

This paper presents a novel probabilistic approach that relies solely on filtered boarding counts and GPS coordinates of boarding to estimate the occupancy level of a bus line within a transportation network. In order to determine occupancy levels, alighting counts are calculated for bus lines where the inward and outward route directions of a bus line run against each other nearly parallel, and the day-end passenger counts of each route are somewhat balanced. The data for the alighting estimation were obtained from the Automatic Fare Collection (AFC) system, which includes both bus stop information and the boarding itself. Video recordings are analyzed to determine actual alighting and compared with the estimated counts for accuracy calculations.

The proposed method, which does not require the utilization of training data, represents a pioneering approach for generating highly accurate estimates on a per-service-line basis by reversing the route. Furthermore, our approach employs line-specific peak hour intervals instead of the traditional fixed rush hour times, thereby enhancing its effectiveness and efficiency. Unlike other methods that track all successive trips of commuters, our approach only requires boarding counts at the target line's stops. This not only reduces the calculation time, but also ensures anonymity.

Using the AFC boarding data, transportation authorities can predict vehicle occupancy levels and choose the optimal vehicle types and sizes for a given route. This enables them to set route trip frequencies dynamically while balancing cost and comfort. As a result, better management of resources leads to the optimization of the transportation network, an increase in ridership, and a more sustainable urban environment.

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.

Acknowledgments

The authors would like to thank Ulaşım Planlama ve Raylı Sistem Dairesi Başkanlığı - Antalya for the data provided in this pilot study.

References

A. Ait-Ali, J. Eliasson, Dynamic origin-destination-matrix estimation using smart card data: an entropy maximization approach, RailNorrköping2019.Norrköping, Sweden (2019).

M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, Knime: The konstanz information miner, In: Studies in Classification, Data Analysis, and Knowledge Organization, 2007.

Bertsimas, D., Sian Ng, Y., Yan, J., 2020. Joint frequency-setting and pricing optimization on multimodal transit networks at scale. Transp. Sci. 54, 839–853.

Bulut, B., Gunay, M., Ozgun, K., Ledet, J., 2021. Optimizing bus lines using genetic algorithm for public transportation. Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. 46, 131–136.

Cheng, Z., Trépanier, M., Sun, L., 2021. Probabilistic model for destination inference and travel pattern mining from smart card data. Transportation 48, 2035–2053.

Faroqi, H., Mesbah, M., Kim, J., 2018. Applications of transit smart cards beyond a fare collection tool: a literature review. Adv. Transp. Stud. 45.

Farzin, J.M., 2008. Constructing an automated bus origin–destination matrix using farecard and global positioning system data in sao paulo, brazil. Transp. Res. Rec. 2072, 30–37.

L. Gannes, Moovit Powers Public Transit App with the Wisdom of the Crowd, Technical Report, Moovit, 2012.AllThingsD.com.

D. Glasgow, Google Maps is turning 15, Technical Report, Google, 2023.⟨https://www.blog.google/products/maps/maps-15th-birthday/⟩.

GlobeNewswire, Global automated passenger counting and information system market report 2022 to 2027: Growing technological developments in automated passenger counting systems presents opportunities, ⟨https://www.globenewswire.com/en/news-release/2022/10/27/2542726/28124/en/Global-Automated-Passenger-Counting-and-Information-System-Market-Report-2022-to-2027-Growing-Technological-Developments-in-Automated-Passenger-Counting-Systems-Presents-Opportunit.html⟩, 2022.Accessed: 2023–07-10.

Harrison, G., Grant-Muller, S.M., Hodgson, F.C., 2020. New and emerging data forms in transportation planning and policy: opportunities and challenges for "track and trace" data. Transp. Res. Part C: Emerg. Technol. 117, 102672.

Hussain, E., Bhaskar, A., Chung, E., 2021. Transit od matrix estimation using smartcard data: recent developments and future research challenges. Transp. Res. Part C: Emerg. Technol. 125, 103044.

Jung, J., Sohn, K., 2017. Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. IET Intell. Transp. Syst. 11, 334–339.

Lee, E., Cen, X., Lo, H.K., Ng, K.F., 2021. Designing zonal-based flexible bus services under stochastic demand. Transp. Sci. 55, 1280–1299.

Li, T., Sun, D., Jing, P., Yang, K., 2018. Smart card data mining of public transport destination: a literature review. Information 9, 18.

Lu, K., Liu, J., Zhou, X., Han, B., 2020. A review of big data applications in urban transit systems. IEEE Trans. Intell. Transp. Syst.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C: Emerg. Technol. 36, 1–12.

X. Ma, Smart card data mining and inference for transit system optimization and performance improvement, Ph.D. thesis, University of Washington, 2013.

Mohammed, M., Oke, J., 2022. Origin-destination inference in public transportation systems: a comprehensive review. Int. J. Transp. Sci. Technol.

Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M., 2011. Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. Transp. Res. Rec. 2263, 140–150.

K. Özgün, M. Günay, B.D. Başaran, Determination of peak times in public transportation, In: 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), IEEE, 2021a, 1–6.

Özgün, K., Günay, M., Başaran, B.D., Bulut, B., Yürüten, E., Baysan, F., Kalemsiz, M., 2021b. Analysis of public transportation for efficiency. In: Hemanth, J., Yigit, T., Patrut, B., Angelopoulou, A. (Eds.), Trends in Data Engineering Methods for Intelligent Systems. Springer International Publishing, Cham, pp. 680–695.

Özgün, K., Başaran, B.D., Günay, M., Ledet, J., 2023. Boarding pattern classification with time series clustering. Smart Applications with Advanced Machine Learning and Human-Centred Problem Design. Springer International Publishing, Cham, pp. 691–699.

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. Part C: Emerg. Technol. 19, 557–568.

Pi, X., Qian, Z.S., Steinfeld, A., Huang, Y., 2018. Understanding human perception of bus fullness: an empirical study of crowdsourced fullness ratings and automatic passenger count data. Transp. Res. Rec. 2672, 475–484.

Redman, L., Friman, M., Gärling, T., Hartig, T., 2013. Quality attributes of public transport that attract car users: a research review. Transp. Policy 25, 119–127. https://doi.org/10.1016/j.tranpol.2012.11.005.

A. Steinfeld, Tiramisu Transit, Technical Report, Carnegie Mellon University, 2023.⟨https://tiramisutransit.com⟩.

Stewart, C., Bertini, R., El-Geneidy, A., Diab, E., 2016. Perspectives on transit: potential benefits of visualizing transit data. Transp. Res. Rec. J. Transp. Res. Board 2544. https://doi.org/10.3141/2544-11.

E.J. Taaffe, Geography of transportation, Morton O'kelly, 1996.

Welch, T.F., Widita, A., 2019. Big data in public transportation: a review of sources and methods. Transp. Rev. 39, 795–818. https://doi.org/10.1080/01441647.2019.1616849.

Yang, Y., Liu, J., Shang, P., Xu, X., Chen, X., 2020. Dynamic origin-destination matrix estimation based on urban rail transit afc data: deep optimization framework with forward passing and backpropagation techniques. J. Adv. Transp. 2020.

Yang, X., Xue, Q., Ding, M., Wu, J., Gao, Z., 2021. Short-term prediction of passenger volume for urban rail systems: a deep learning approach based on smart-card data. Int. J. Prod. Econ. 231, 107920.

Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T., 2019. Big data analytics in intelligent transportation systems: a survey. IEEE Trans. Intell. Transp. Syst. 20, 383–398.