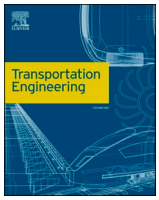
Transportation Engineering 16 (2024) 100239

Contents lists available at ScienceDirect

Transportation Engineering

journal homepage: www.sciencedirect.com/journal/transportation-engineering



Moving from classical towards machine learning stances for bus passengers' alighting estimation: A comparison of state-of-the-art approaches in the city of Lisbon

Sofia Cerqueira [a,b,*], Elisabete Arsenio [a], José Barateiro [c], Rui Henriques [b]

[a] Institution: National Laboratory of Civil Engineering, Lisbon, Portugal
[b] INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
[c] Faculty of Sciences and Technology, University of Algarve, Faro, Portugal

ARTICLE INFO

Keywords:
Alighting estimation
Trip-chaining
Density-based clustering
Non-commuting patterns

ABSTRACT

Passenger alighting estimation is a critical task in public transport (PT) management, especially for entry-only Automatic Fare Collection (AFC) transport systems where passenger alighting are not recorded. Effective estimation methods are necessary for trip analysis and route planning, offering valuable insights into passengers' mobility patterns and, subsequently, improving the quality of service. However, the stochastic nature of passenger behaviour challenges the degree of successful alighting estimates. A classic approach to infer the alighting stops of passengers is the use of trip-chaining principles. Since these principles are dispersed across the literature in the field, their comprehensive review is pivotal to establish the best practice for alighting estimation. Still, trip-chaining approaches are unable to infer the alighting of non-commuting passengers. This paper addresses these two research gaps by: i) providing a critical overview of the existing principles and methods for alighting estimation; ii) proposing an approach to improve alighting estimation that consistently integrates the most effective state-of-the-art principles on trip-chaining; and iii) further introducing a frequent pattern mining and density-based clustering solutions to support alighting estimation for non-commuting passengers. Considering the public bus transport in Lisbon city as the guiding case study, the achieved estimation rate by the proposed assembled model is 92%. Moreover, the density-based clustering solution is found to improve the estimation of 11pp against classic trip-chaining principles. Furthermore, the proposed model and acquired results yield actionable value to enhance PT operations and services, ultimately leading to improved bus routing and quality of service.

1. Introduction

Passenger alighting estimation in entry-only AFC systems plays a leading role in PT planning and urban management [1,2]. Boarding and alighting data are essential to provide a roadmap for sustainable transport and mobility in cities, e.g., enhancing the accessibility and space of senior citizens to promote social inclusion [3,4] and ensuring efficient access to goods and services for the city's economic growth. With the goal of promoting full coverage of passengers' alighting, different authors have proposed principles (or assumptions) as heuristics to conduct the alighting estimation. As a result, the assessment of existing state-of-the-art principles is dispersed across different studies which use various methods and assumptions in each context. Therefore, the impact of the most useful principles must be quantified, along with a critical analysis of the relevant literature. Furthermore, the stochasticity of passengers' behaviour and multimodal transport choices pose challenges for the alighting estimation since these principles can only describe general commuting behaviour. To address the above issues, this research aims to improve the effectiveness of bus alighting estimation, using smart card data from Lisbon public operators, buses, and subway. To this end, the paper contributes with i) a critical point of view on the state-of-the-art principles, identifying its limitations and research opportunities; ii) extended principles to explain commuter travelling; iii) a three-stage alighting model that considers not only the best practices (principles) but as well user-centric pattern mining; and iv) data profiling trips with unsuccessful estimation by exploring discrepancies

* Corresponding author.
E-mail addresses: scerqueira@lnec.pt, sofia.cerqueira@tecnico.ulisboa.pt (S. Cerqueira), elisabete.arsenio@lnec.pt (E. Arsenio), jebarateiro@ualg.pt (J. Barateiro), rmch@tecnico.ulisboa.pt (R. Henriques).

https://doi.org/10.1016/j.treng.2024.100239
Received 28 April 2023; Accepted 3 March 2024
Available online 9 March 2024

in card titles, schedules and in the city.

The paper is organized as follows: Section 2 discusses the opportunities and limitations of state-of-the-art principles and describes the latest methods for PT alighting inference; Section 3 proposes a three-stage model that integrates trip-chaining principles, density-based clustering and frequent pattern mining; Section 4 gathers the main results of each estimation stage and carries an exploratory analysis of trips without alighting data, as a final validation of the results; and finally, Section 5 provides final remarks.

2. Literature review

This section discusses limitations and future opportunities of state-of-the-art principles for alighting estimation in entry-only AFC systems, along with recent methods for inference based on probabilistic and machine learning stances.

2.1. Principles: opportunities and limitations

Entry-only AFC systems are prevalent in worldwide transport systems, particularly in bus operators. In this context, alighting estimation has been largely pursued as a primary step for operators to understand the final destinations of their passengers and the existing centre-flow patterns [5–7]. Trip-chaining algorithms are commonly used to infer alighting data using smart card data. However, we observe significant differences pertaining to the selection and implementation of the trip-chaining principles [6,8–12]. A principle describes the general behaviour of passengers in certain conditions. Considering the trip-chaining approach, a principle can be transformed into a heuristic that determines which trip should be chained with the target trip. The alighting stop of the target trip is estimated by minimizing the distance or time spent between the two trip segments. Moreover, a threshold determines if the alighting stops are inferred or not. In the literature, these parameters range up to 2000m. Most authors define the threshold between 600 and 1000m to avoid overestimation [10,11,13]. This definition generally depends on several factors such as walking habits, network structure and geological aspects of the city.

To complement the analysis of Table 1, the state-of-the-art principles for alighting estimation in entry-only AFC systems are described. Furthermore, the strengths, possible limitations, and opportunities for improvement each principle are discussed. Moreover, the principles' tag describes its purpose. S or E or V tags indicate that the principle is responsible for approving a trip (A) or estimating (E) or validating (V) the inferred alighting stop. A principle whose tag starts with O means that is not implemented in the trip-chaining approach.

A1: For all passengers, a day starts when the network has the lowest activity, for instance, from 4 AM to 3:59 AM of the next day [6,14].
A2: Two consecutive boarding stops will not occur in the same location [8].

Given that the trip-chaining approach needs as input a set of valid trips, of a given day, ordered by boarding time, principles A1 and A2 are responsible for validating whether a trip must belong to the estimation process. Principle A1 resulted from Munizaga et al. [6] validation of inference results. The author's analysis shows that unsuccessful inferences are more prone to appear on trips that occur at the end of the day. The issue was successfully solved by shifting 24 h period from midnight to 4 AM (the hour of less activity on the network), ending at 3:59 AM. Principle A2 suggested by Barry et al. [14] avoids the boarding of two consecutive trips occurring in the same location, e.g., the passenger accidentally taps two times. In this case, the first trip is not valid for estimation.

E1: After alighting, the passenger will walk to the next boarding, whose transfer distance must be less than a given threshold (except for the last trip of the day) [8].
E2: For the last trip of the day, the passenger will alight close to the first boarding of the day, whose distance must be less than a given threshold [9].

As shown in Table 1, principles E1 and E2 are widely used in trip-chaining algorithms and have been validated in several research scenarios [6–26]. According to studies, these principles can explain the behaviour of frequent and commuting passengers. Given the complexity of travel patterns and the uniqueness of certain public networks, other principles were proposed to improve the estimation rate, particularly the inference of the last trip's alighting stop.

E3: For the last trip of the day, if E2 does not work then, the alight stop can be estimated by assuming that is close to the first boarding of the next day [9].
E4: If the model is unable to successfully infer an alighting stop, a candidate alighting stop can be among the stops on the same route, but in the opposite direction [12].

To enhance the last trip's alighting inference, Trépanier et al. [9] suggest relaxing the assumption E2 by articulating the idea in E3. However, the author does not take into account the fact that these two consecutive boarding stops may occur at the same or in a nearby area, e.g., a residence. In this case, the first stop's alighting stop cannot be estimated. Considering this issue, our previous work suggests improving the principle along with a new parameterization [26]. First, the new principle considers chaining the last trip of the day with the next boarding stop of the following days (not just the next day). Second, consecutive boarding stops must be at least a given distance.

Behind the principle E4, Nassir et al. [12] reasonably assume that the GPS from the AFC system misidentifies the boarding stop location. This issue occurs often when the same route in different directions shares the same terminal. The author does not specify, though, if the stations on the entire route are being searched for, or only those upstream from the boarding stop across the street. In fact, considering all route' stops may lead to incorrect inference.

V1: A candidate alighting stop must have an arrival time no later than the boarding time of the next trip [11,12].
V2: Passengers only alight in the allowed fare zones [10].
V3: Passenger taps their card multiple times during a trip in order to see the time left to travel [10].

The principles V1 to V3 are responsible for validating the candidate alighting stop of all passengers' trips. Nassir et al. [12] propose principle V1 that considers invalid a candidate alighting stops if the arrival time is later than the boarding time of the consecutive trip. Using the same idea, Alsger et al. [11] only exclude an alighting stop if the arrival time is later than the next boarding time plus a few minutes. The majority of studies use expected timetables to determine arrival times. However, in most cases, the expected arrival time does not take into account factors that affect traffic, such as the weather, the existence of big events, and road interruptions. Therefore, using principle V1 as Alsger et al. [11] suggest may not produce the desired results. Having access to the actual arrival time of the candidate's alighting stop would be a preferable approach. Finally, Nunes et al. [10] suggest principles V2 and V3. Principle V2 states that a candidate alighting stop turns valid if it belongs to the allowed travel. While V3 validates a candidate alighting stop if its sequence number is higher than the last stop sequence recorded by the passenger. These principles are recommended as good practice, however these are only applicable in research scenarios that have bounding zones and the passenger can tap more than once.

Table 1
Mode *B* = Bus, *S*=subway, *R*=Rail, *T*=Tramways, *F*=Ferry; E1-E4, A1, A2, V1-V3, O1, O2 = Principles; ML = Machine learning inference methods; ER*= Estimate rate according to the criteria of the respective author; Final Study Purpose IAE = Improve Alighting Estimation, OD= Origin- Destination Inference.

Ref and Study case location	Mode					Dataset	Principles											ML	Validation Method for alighting estimation	ER*	Final Study Purpose		
	B	S	R	T	F		E1	E2	E3	E4	A1	A2	V1	V2	V3	O1	O2				IAE	OD	Others
(Trépanier et al., 2007) Gatineau, Canadá	X					July and October 2013	X	X	X		X								No validation	66 %	X		
(Zhao et al., 2007) Chicago, USA	X		X			6 days, January 2004	X	X								X			No validation	64 %		X	
(Farzin et al., 2008) São Paulo, Brasil	X					May 9, 2006	X	X											Comparative analysis between inferred ODs with OD Household Surveys	NA		X	
(Barry et al., 2009) New York City, USA	X	X		X	X	2 weeks of April 2014	X	X	X					X			X		Ride check data with the counting of a couple of routes is used as validate.	90 %	X	X	
(Nassir et al., 2011) Minneapolis, EUA	X					November 10, 2008	X	X		X	X		X						Sensitive analysis on model parameterization	60.7 %	X	X	
(Li et al., 2011) Jinan, China	X					Route 115	X	X											No validation	70 %		X	
(Munizaga et al., 2014) Santiago, Chile	X	X				1 week	X	X	X		X								First, an exploration analysis on the model results. Second, comparison of inference results with trips provided by volunteers	84.2 %	X		Route choice, and trip purpose
(Alsger et al., 2015) SEQ Australia	X		X		X	1 week	X	X											Sensitive analysis on model parameterization to test assumptions	NA	X	X	
(He et al., 2015) Gatineau, Canadá	X					October 2009	X	X	X				X			X		X	No validation	91.54 %	X		
(Nunes et al., 2015) Porto, Portugal	X					April 2010	X	X			X			X	X				Principles V2 and V3 are used to validate the inferred alighting stops	62.4 %	X	X	
(Alsger et al.,2016) SEQ, Australia	X		X		X	1 week, 2013	X	X	X				X						Comparative analysis between inferred ODs and real ODs data	72.6 %	X	X	
(Hora et al., 2017) Porto, Portugal	X					January 2013	X	X			X								No validation	NA		X	
(Jung and Sohn, 2017)	X					Single day	Not applicable											X	Comparative analysis with real alighting data.	87 %	X		
(Zou et al. 2018) Beijing, China	X	X				1 week	X	X											No validation	NA			Detect of home location and trip purpose
(Yan et al., 2019) Beijing, China	X	X				Oct 8th to 14th 2016	X	X	X				X					X	Comparative analysis with real alighting data.	70.2 %-74.4 %	X		
(Assemi et al., 2020) SEQ Australia	X		X		X	20 March 2013	X	X		X						X		X	Comparative analysis with real alighting data.	79.5 %	X		
(Liu et al. 2021) Michigan, EUA	X					October 2017	X	X											No validation	NA		X	
(Lee et al., 2021) Seul, Korea	X	X				April 2020	X	X								X			Sensitive analysis on the model by varying transfer distance	85 %	X		
(Lei et al., 2021) Nanjing, China	X	X				July 2020	X	X								X		X	Sensitive analysis on model parameterization	NA	X	X	

O1: Assuming that the passenger has routines, it is likely the passenger takes trips with similar boarding stop, times and routes [9,16,17].
O2: A passenger with a trip that cannot be chained (for instance a single trip), is assigned an alighting stop by random sampling the distribution from other passengers whose boarding stops are the same as the current trip [14].

When a trip is not chainable with others or the trip-chaining algorithm does not estimate with success, O1 and O2 can be applied in a second approach. Some authors consider principle O1, which states that passengers may exhibit patterns along their trips' history. Trépanier et al. [9] propose to look for symmetric patterns to estimate an alighting stop for trips that belong to the pattern. For example, given the patterns, R to B and B to R, a trip with boarding in R is likely to have an alighting stop in B. Subsequent research, including He and Trépanier [17] and Lee et al. [23] followed the same concepts but used different implementation strategies. Finally, principle O2 was not followed by any former works and its effectiveness was not clearly validated by the primary study.

Considering a trip-chaining algorithm, Fig. 1 depicts a passenger's sequence of trips during a day to illustrate the application of principles. Except for the last trip of the day, principle E4 is only applicable when E1 fails in the estimation. Meanwhile, for the final trip of the day, E2 is obligatorily the first principle used, followed by E3 and E4, or vice versa. Considering the order of these principles, the estimation for a given trip ends when one of the principles successfully estimates an alighting stop. The other principles A1, A2, V1, V2 and V3 are also present in the approach to validate all trips and candidate alighting stops.

As shown in Table 1, in cities with more than one transport mode, incorporating multiple modes aids the alighting inference process. Barry et al. [14], Munizaga et al. [6], Alsger et al. [11] and Cerqueira et al. [27] are notable examples that integrated other modes besides the bus, e.g., the subway. In these cases, the estimation rate, defined as the number of trips with an alighting stop estimated divided by the total of trips, is superior. However, these percentages should be considered cautiously, because different study-specific factors can impact results, such as the threshold, the presence of multiple modes, the guiding principles, and the availability of additional approaches to enhance the estimation rate.

Concerning validation methods, alighting estimation in most case scenarios is an unsupervised learning process. Validation options in literature and practice can include: i) comparative analysis against survey data; ii) comparing the results of trip-chaining with volunteers' mobility data; and iii) a sensitive analysis by varying model parameterization [6,11,12,23]. The first option is less effective, while the second is impractical since it requires volunteers. Consequently, among the author's options, a third is preferred. This method involves quantifying the impact of model parameterization, e.g., transfer distance, evaluating the effects of each principle, and providing an exploratory analysis of the model's results.

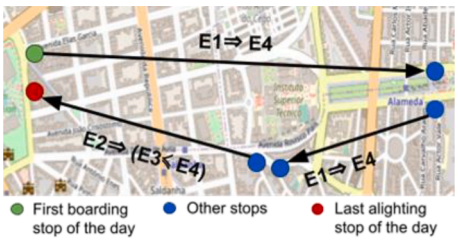


Fig. 1. Illustration of a sequence of trips during a day made by a passenger and the respective applicable principles.

2.2. Machine learning advances for alighting stop estimation methods

Due to the inadequacy of the trip-chaining algorithm for estimating the alighting of irregular and non-commuting trips, recent studies have proposed alternative approaches. The majority proposes machine learning or probabilistic approaches to complement the trip-chaining algorithm [13,21,23]. Yan et al. [13] propose machine learning models to complete the standard method, including Naive Bayesian, Support Vector Machine, Decision Tree, K-nearest neighbours and an Ensemble learning model. First, irregular users, e.g., tourists and occasional PT users, are separated from regular users through the clustering method. Second, both data groups are subject to two-step inference, a trip-chaining algorithm and then one of the machine learning methods stated. To perform training and testing of the machine learning model, the author uses cross validation. Likewise, Assemi et al. [21] propose a trip-chaining algorithm as the first step and then a neural network for irregular trips. Jung and Sohn [24] proposed a deep-learning model solely based on both smart-card data and land-use features. Despite the advances over the baseline model, the author applies rigid assumptions that may lead to overestimation. For instance, Lee et al. [23] assume that for a trip that occurs before midday, the boarding stop is more likely to be near the passenger' home. However, this principle conflicts with the unpredictable nature of passengers' behaviour on PT, e.g., a commuter may leave to work in the afternoon and arrive home at night. Lei et al. [25] propose a framework for enhancing alighting estimation in single trips, based on the Minimum Entropy Rate criterion. In the context of public transportation, a single trip refers to a unique passenger's trip made within a day. The approach uses the principles of noise reduction in information theory. Aiming to preserve travel regularity in passenger mobility sequences, a potential alighting stop is determined by minimizing the entropy rate.

He and Trépanier [17] follow up Trépanier et al. [9] by suggesting a kernel density estimator to improve the existing trip-chaining algorithm. The method jointly models spatial (discrete variable) and temporal (continuous variable) probabilities to determine the likelihood of each stop candidate. Despite the relevance of the surveyed advances on unsupervised alighting inference, He and Trépanier [17] only take into account the spatial probability of alighting stops, whereas boarding stops should be also considered. Furthermore, the distance between trip segments is critical in assessing the validity of the inferred alighting stop, which remains largely unexplored on the aforementioned probabilistic approaches.

3. Three-stage model and validation for alighting estimation

Considering the surveyed advances, we propose an ensembled model for alighting estimation based on three major stages (described in Sections 3.1 to 3.3). The hyperparameterization and validation methods are described in Sections 3.3 and 3.4, respectively.

3.1. First stage: trip-chaining principles

In light of the principles mentioned in Section 2, principles E1 and E2 were chosen as a baseline to conduct the trip-chaining algorithm, due to the ability to generalise commuting behaviour. As previously discussed, E3 and E4 have limitations, so the proposed algorithm includes extended versions of these, called B1 and B2, respectively. Moreover, A1 and A2 are here used to validate the input trips. Meanwhile, V1, V2 and V3 are not considered in the validation of the alighting stops, since V1 may induce more inference errors and V2 and V3 are not applicable to our research scenario. For simplicity, let a trip be a single travel from one point to another using only one vehicle of public transportation.

Considering an ordered set of trips for a given passenger, principle B1 states that an alighting stop for the last trip of the day (target trip) can be estimated by assuming that it is close to the first boarding that occurs in the next days (rather than the next day only, as states principle E2). B1 is

only applicable if principle E2 does not estimate an alighting stop before. Considering the estimation based on Principle B1, there is a chance that two stops on successive boarding trips will be near to one another, for instance, the passenger's home. Therefore, B1 is only feasible if the consecutive boarding stops are at a distance higher than a given threshold. For simplicity, the threshold is called NCB (no close boarding). Principle B2 states that a candidate alighting stop can be among the stops upstream from the stop across the street, in the opposite route (rather than considering all stops from the opposite route). B2 is only applicable when none of the previous principles performs the estimation successfully.

The trip-chaining algorithm receives the daily bus and subway trips of each passenger as input. Consistently with principle A1 and the public network activity of our study case, a day starts at 3:59 AM until 4 AM the next day. For additional details, we refer to Cerqueira et al. [27] for details on the consistent integration of trip-chaining principles [26].

3.2. Second stage: density-based clustering

Clustering methods are frequently used to unravel spatial and temporal patterns in public traffic flow, as well as support other relevant tasks including anomaly detection, trip purpose profiling, segment passenger, among others [8,23,28–30]. This research proposes identifying the sequence of main clusters of each passenger, to support the alighting estimation. Using density-based clustering, GPS passenger locations, alighting and boarding coordinates are clustered. In these terms, the principle O1 of the literature review, which states: "Assuming the passenger has routines, it is likely the passenger takes trips with similar boarding stop, times, and routes," is the underlying principle of estimation in this approach. Furthermore, a temporal criterion limits the number of potential clusters that can be used to search for a candidate alighting stop. Consider the following example to demonstrate this temporal criterion: if a passenger traveled from cluster A to cluster B (or from cluster B to A) within a single day, then cluster B is taken into consideration as a potential cluster. Considering a set of trips for a given passenger:

1. DBSCAN clustering is applied to cluster all boarding and alighting stops.
2. For each trip, the boarding and the alighting cluster identifier are assigned.
3. A sequence of boarding and alighting clusters for each day and ordered temporally is generated as $C_d = \{c_1, c_2, \cdots c_j\}$, where d denotes a day. The sequence also includes missing alighting clusters with a unique identifier.
4. Given a trip t whose alighting stop and cluster are missing, the respective boarding cluster b is gathered.
5. In the sequences C_d, cluster b is searched and its neighbours are gathered. In other words, if $\exists(c_j = b \mid c_j \in C_d)$, then c_{j-1} and c_{j+1} are saved in a list of potential clusters P.
6. The distance to each cluster on P for each candidate alighting stop of t is calculated. The centroid of the cluster the stop with the most frequent boardings or/and alightings.
7. The alighting stop of t is determined by the shortest distance between a candidate stop and a potential $d = (Min(D(a,p) \mid a \in \mathrm{A},\ \mathrm{p} \in \mathrm{P}))$, where A is a set of candidate alighting stops that are upstream of the boarding stop t. If $d > T$, where T is a defined threshold, t remains without alighting stop.

3.3. Third stage: frequent patterns

As the final stage, the model incorporates principle O1, mentioned in Section 2. Considering a target trip without an alighting stop, trips with the same boarding stop and similar timestamps are gathered from the passenger's sample. An acceptable alighting stop is chosen from this subset if two conditions are met: i) the candidate stop occurs more frequently than a predefined threshold, and (ii) the boarding time difference between the trips and the target trip is less than the given period. This principle resembles and emulates the behavior of frequent pattern mining stances for alighting estimation [9,16,17].

3.4. Three-stage model hiperparameterization

Considering the proposed ensemble model, this section explains the parameterization of each stage. Starting with the first stage, the presented herein trip-chaining model has two parameters, which are: i) the maximum allowed transfer distance, and ii) the distance between consecutive boardings (NCB). Both parameters are responsible for validating whether candidate alighting stops are admissible or not. In our previous research, these parameters were submitted to a sensitivity analysis [26]. The maximum transfer distance and NCB values chosen were 1000 and 2000m, respectively. The values were selected to avoid overestimation.

The second stage implies the parameterization of: i) the density-based clustering (distance thresholds ε and neighbourhood size κ), and ii) the distance between the candidate stop to the target cluster. Considering density-based clustering parameters, a cluster is valid if it contains one or more instances, so κ is set to 1. Notice, let be an instance an alighting or boarding stop coordinates from passenger trip. Meanwhile, training and testing were carried out to determine the best ε value. For this purpose, the initial training set-up consists of a dataset with 1500 passengers and distance variation ε from 100m to 1500m, with a step of 100m. During the testing phase, the Silhouette coefficient and Calinski Harabasz score were the metrics used to quantify the distance between clusters and cluster cohesion. According to the results, 600m is the best eps for the majority of passengers. The third parameter is the distance between the candidate stop to the target cluster. In reality, this distance corresponds to the maximum allowed transfer distance present in the first stage. Hence, the third parameter is set as well to 1000 metres.

Finally, for the third stage of the model, the chosen parameterization is according to Lee et al. [23] suggestions, i.e., an alighting stop is only inferred if the stop occurs more frequently than three times in a week. Furthermore, the boarding time difference between the trips and the target trip is less than a 2 h period. The strict parametrization here practised aims to prevent the overestimation, and consequently provide more reliable results on estimation rate.

3.5. Validation

Along with the model hiperparameterization, several steps were taken to meet the potential of each stage. Additionally, other validation methods are employed: i) a sensitivity analysis, ii) a comparative analysis between two models, and iii) an exploration analysis on trips without alighting data.

First, a sensitivity analysis is provided by analysing the impact of each estimation phase. Particularly the impact of principles E1, E2, B1 and B2 in the first stage are analysed in detail. Notice, since principles A1 and A2 are heuristics responsible for the validate the input trips, their assessment is not considered. To quantify the effectiveness of each stage, the following metrics are provided: i) the estimation rate, defined as the number of trips with alighting stops estimated divided by the total of the smart card data trips used; ii) statistical analysis of transfer distance; and iii) average value of confidence score associated to each alighting stop estimated. According to each stage, different metrics are applied to obtain the confidence score of each alighting stop estimated. For the first and second stages, the metric applied is $Cd1(w)$ where w is the distance between the estimated to next boarding stop segments (stage 1) or distance between the estimated alighting stop and potential cluster (stage 2), i.e. transfer distance. Meanwhile, for the third stage, the metric applied is $Cf(\varphi)$, where φ is the frequency of a pattern, i.e. the number of

trips with the same boarding and alighting stop,

$$Cd1(w) = \begin{cases} 100, & if\ w \leq minT \\ \dfrac{maxT - w}{maxT - minT} * 100, & if\ minT < w \leq maxT \\ 0, & otherwise \end{cases} \quad (1)$$

$$Cf(\varphi) = \begin{cases} 100, & if\ w \leq minF \\ \dfrac{\varphi - \mathrm{minF}}{maxF - minF} * 100, & if\ minF < \varphi \leq maxF \\ 0, & otherwise \end{cases} \quad (2)$$

where [*minT, maxT*] and *[minF, maxF*] are the adopted ranges for the transfer distance and pattern frequency values. The assigned score ranges between 0 and 100 where a high value indicates high confidence in the estimation. Based on the reasons mentioned in the previous section, the assigned default values are 200 and 1000m for *minT* and *maxT*. Grounded on Lee et al. work [23], minF and *maxF* variables are set to 12 and 15 (minimum and maximum values of 3 times per week), respectively.

The model is trained using both bus and subway data. Additionally, a separate model is trained using only bus data. The purpose of this approach is to quantify the match between the two models, and better assert the contribution of each stage to the estimation process.

Finally, an exploratory analysis is further conducted on trips without alighting data. This analysis aims at identifying vulnerable profiles, schedules and stops for imputation. In summary, by training and comparing two different models and conducting an exploratory analysis, this study provides valuable insights into the estimation process and identifies potential areas for improvement.

4. Results and discussion

Smart card data from public transport in Lisbon is used as the guiding study case. In detail, the available datasets came from the public bus and metro transport operators, named CARRIS and METRO, respectively. Overall, estimation methods are validated on a complete monthly sample with over 44 million trips. From this count, 11 million trips are from the public bus operator, whose alighting information is missing.

4.1. Overall results

As shown in Table 2, the estimation rate, considering all three stages, is 92.84%. In the first stage, the corresponding results show the performance rate of principles E1, E2, B1, and B2. In total, the approach yields a 79.44% estimation rate. And, as observed in related studies, principles E1 and E2 contribute to the majority of the estimation rate. In the second stage, the density-based clustering approach can significantly increase the estimation rate in 10.93pp. Furthermore, the second stage enables to infer the alighting stops for a single trip, corresponding to 5.13pp. In the third stage, frequent trips with the same boarding and similar boarding timestamp residually increase the estimation rate by 0.17pp. Considering the first and second stages, an analysis of averages and quartiles demonstrate that the transfer distance ranges up to the maximum value of 240m. These results are summarised in the confidence metric based on the transfer distance whose range is between 91.4 to 97.1%. Concerning the third stage, the revealed confidence metric based on the frequency of the pattern is 59.6%. Overall, the results demonstrate the model's effectiveness in alighting estimation.

As the second method of validation, the same estimation model was employed with only data from the bus operator. For simplicity, the former model can be referred to as multimodal and this one as unimodal. As expected, the estimation rate is lower than the multimodal model outcome, corresponding to 83%. Comparing the output of the two models, we find a match of 73.32pp at 83.0%. A more in-depth examination revealed that 9.5pp of the match was estimated by stage 1 in the multimodal model and by stages 2 and 3 in the unimodal model. This shows that stages 2 and 3, in the absence of multimode data, can produce the same outcomes as the trip-chaining in a multimodal model.

4.2. Analysis by card profiles

Fig. 2 depicts the difference in the estimated percentage of trips without alighting stops between the first stage and the following stages (second and third), against the total trips of each card title group. Because the third stage's estimation rate is residual, the evaluation is grouped with the second stage. The chosen card titles are the eighth most representative of the entire sample. Considering the final outcome from each card title, *Zapping* and *Bilhete 24 Horas CA/ML Rede* are the groups with the highest percentage of trips without an inferred alighting stop. In fact, except for these two titles, the remaining are monthly fare cards. Occasional travellers, for instance, tourists, tend to have irregular or/and non-frequent travel patterns. Unlike commuters, tickets with the lowest durability are the first choice for this profile.

Furthermore, encouraging outcomes are discovered for the remaining card titles used by commuters. Among the cards mentioned, three with *Nav. Metropolitano* in the name description have a significant decrease between the first and subsequent stages, corresponding to 15/16pt. The *Nav. Metropolitano - Normal, Nav. Metropolitano - Social+*, and *Nav. Metropolitano - 4_28/sub 23* cards are designated for adults, financially vulnerable passengers, and young people under the age of 23, respectively. Regardless of the card's title, *Nav. Metropolitano* is the only monthly fare card that allows unlimited trips between different public transportation operators and municipalities in the Lisbon district. As a result, the collected data provide preliminary evidence in support of the ability of stages 2 and 3 to generalise in the absence of multimodal data.

4.3. Time-sensitive analysis

Fig. 3 shows the percentage of trips without alighting stops estimated

Table 2
Estimation rate for each phase of the estimation process.

Stage		ER (%)	CER (%)	Transfer Distance (meters)				χ Confidence (%)	
				χ	Q1	Q2	Q3	Distance based	Frequency based
1. Trip-chaining Algorithm	E1	54.44	–	148	26.2	80.5	171.3	95.1	–
	E2	19.62	74.06	198	23.4	74.8	240.1	92.1	–
	B1	2.16	76.22	197	20.5	74.2	212.5	91.4	–
	B2	3.22	79.44	145	22.2	76.9	175.4	94.4	–
2. Clustering	Single trips	5.13	84.57	86	0	21.8	102.2	97.1	–
	Others	5.80	90.37	105	0	31.9	120.1	96.2	–
3. Frequent Pattern		0,17	90.54	–	–	–	–	–	59.6
Not Estimated		9.46	100	–	–	–	–	–	–

*ER(%) = Estimation Rate; CER(%)=Cumulative Estimation Rate; χ= Average; Q1 to Q3= Quartiles.

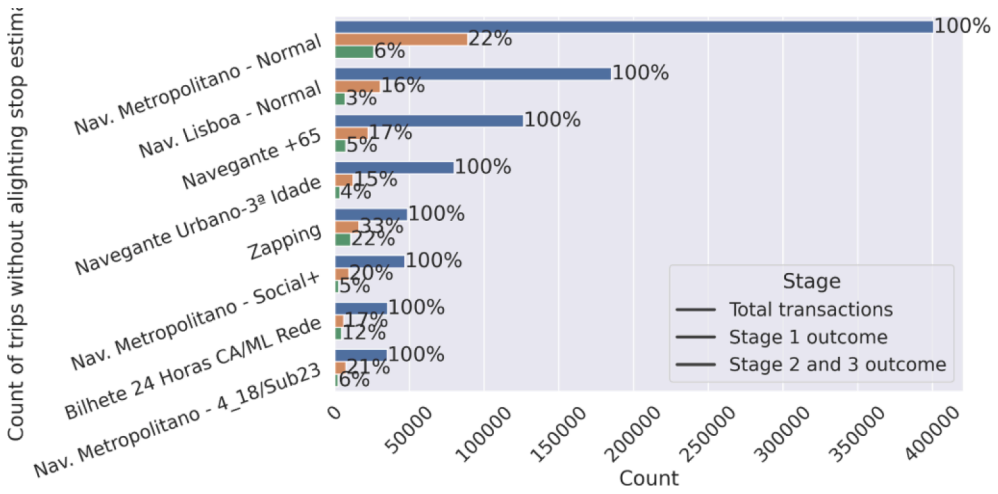


Fig. 2. The difference in the estimated percentage of trips without alighting stops between the first stage and the following stages against the total trips of each card title group.

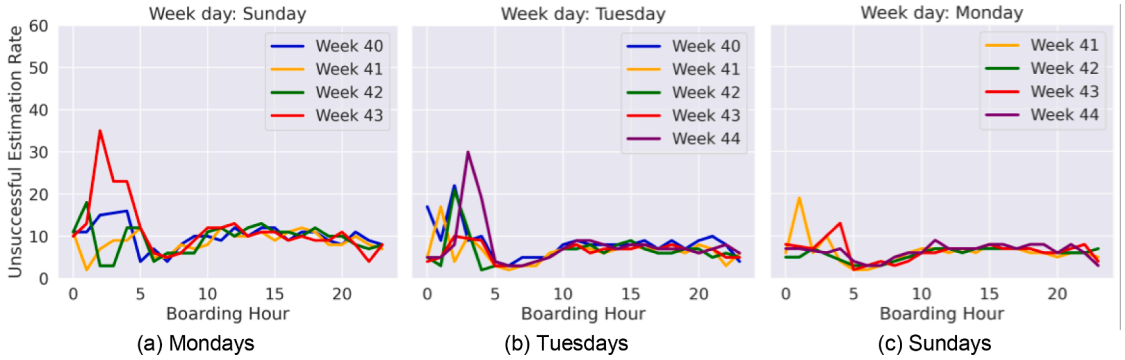
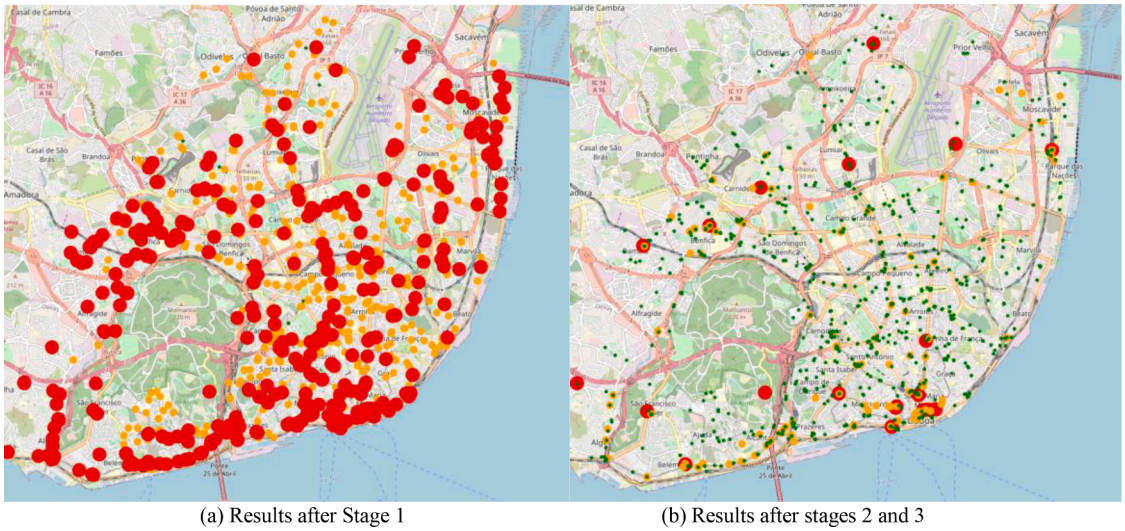


Fig. 3. Distribution of trips without alighting stops estimated by the hour, on Mondays, Tuesdays and Sundays from October 2019.

by hour, throughout the day. Each line on each chart corresponds to a weekday in a given week of the year. All charts highlight that estimation incapacity is significantly noticed between 12 PM to 5 AM, for all days. This disparity appears to indicate that trips of passengers' journeys are not fully contained in the operator's cycle period, that is, from 3:59 AM until 4 AM of the next day.

4.4. Spatial analysis

Fig. 4 depicts the spatial distribution of trips without alighting stop (unsuccessful estimation). The red circles indicate that the alighting stop was not estimated for more than 20% of trips that began at that boarding stop. Orange circles represent a range of 10 to 19%, while green circles



(a) Results after Stage 1 (b) Results after stages 2 and 3

Fig. 4. Spatial distribution of trips without alighting stops estimated, considering the boarding stop location.

represent less than 9% of trips that began at the corresponding boarding stop and the alighting stop was not estimated. According to the aforementioned criterion, Fig. 4a displays the results after stage 1 and Fig. 4b after stages 2 and 3. Observing Fig. 4a, the red and orange circles are equally distributed along the maps, not suggesting areas with a lower rate of alighting estimation. On the other hand, Fig. 4b only shows the red circles in the Lisbon city boundary and the city's historic center in, depicting a significant difference from the previous map. Indeed, these boarding stops (red circles) coincide with transfers between other public operators.

5. Conclusion

This research enhances the alighting stop estimation by contributing to an ensembled model that incorporates two main aspects: i) the integration of dispersed state-of-the-art trip-chaining principles and ii) a clustering approach that gathers meaningful structure of passenger trips. In particular, spatial and temporal relations are gathered to support the estimation by applying density-based clustering on GPS coordinates from all origins and destinations of the passenger. Considering as case study the public transport smart card data in the Lisbon city, the proposed assembled model yields a 92% estimation rate11pp higher in comparison with applying classic trip-chaining principles. Since the previous model receives as input multimodal data (bus and subway), a comparative analysis against a model that uses bus data only (unimodal model) is provided. Comparing the estimates of the previous multimodal setting (bus and subway) with the unimodal setting (bus), we observe alighting concordance in approximately 83% of the estimated cases. A closer look indicated that the clustering approach in the unimodal model, as well as the trip-chaining method in the multimodal model, accurately estimated 9.5 points of the match. These results demonstrate the potential of our machine learning approach in absence of multimodal data.

Along with this endeavour, the work offers, first, a comprehensive analysis of the state-of-the-art principles; second, a bespoken model that considers user-centric pattern mining methods and multimodal traffic data; and third, a significant improvement of the estimation rate, compared with our previous work [26]. Besides, this research undertakes a spatial-temporal analysis on trips with unsuccessful estimation, to further inspect their nature. Aiming to assess the limitations of the model and validate results, a deeper exploration of the smart card data sample is performed by card title, time of the day, and location in the city. Among all findings, the spatial analysis within the Lisbon city PT reveals that boarding stops located along the periphery and/or acting as interface areas modes are the most difficult ones to estimate.

Overall, the highlighted research efforts on alighting stop estimation are novel and of key importance for subsequent ends, including the inference of multimodal origin-destination matrices.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Sofia Cerqueira reports financial support was provided by Foundation for Science and Technology.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors thank the support of CARRIS, METRO and Câmara Municipal de Lisboa, in particular the attention provided by the Gabinete de Mobilidade and Centro de Operações Integrado. This work is further supported by national funds through Fundação para a Ciência e Tecnologia (FCT) under project "Integrative Learning from Urban Data and Situational Context for City Mobility Optimization" ILU (DSAIPA/DS/0111/2018), INESC-ID pluriannual (UIDB/50021/2020) and a FCT research grant (2022.13483.BD).

References

[1] E. Hussain, A. Bhaskar, E. Chung, Transit od matrix estimation using smartcard data: recent developments and future research challenges, Transport Res. Part C: Emerg. Technol. 125 (2021) 103044.
[2] T. Li, D. Sun, P. Jing, K. Yang, Smart card data mining of public transport destination: a literature review, Information 9 (2018) 18.
[3] L. Abdelfattah, D. Deponte, G. Fossa, The 15-minute city: interpreting the model to bring out urban resiliencies, Transportation Res. Procedia 60 (2022) 330–337.
[4] X. Zhao, Y. Ke, J. Zuo, W. Xiong, P. Wu, Evaluation of sustainable transport research in 2000–2019, J Clean Prod 256 (2020) 120404.
[5] S. Cerqueira, E. Arsenio, R. Henriques, Inference of differential origin-destination matrices to assess the spatio-temporal attractiveness of public transport in relation to car travel: a case study in the city of lisbon. European Transport Conference, 2022.
[6] M. Munizaga, F. Devillaine, C. Navarrete, D. Silva, Validating travel behavior estimated from smartcard data, Transportation Research Part C: Emerging Technologies 44 (2014) 70–79.
[7] Q. Zou, X. Yao, P. Zhao, H. Wei, H. Ren, Detecting home location and trip purposes for cardholders by mining smart card transaction data in beijing subway, Transportation (Amst) 45 (2018) 919–944.
[8] H. Faroqi, M. Mesbah, Inferring trip purpose by clustering sequences of smart card records, Transportation Research Part C: Emerging Technologies 127 (2021) 103131.
[9] M. Trépanier, N. Tranchant, R. Chapleau, Individual trip destination estimation in a transit smart card automated fare collection system, J. Int. Trans. Syst. 11 (2007) 1–14.
[10] A.A. Nunes, T.G. Dias, J.F. e Cunha, Passenger journey destination estimation from automated fare collection system data using spatial validation, IEEE Trans. Intell. Transp. Syst. 17 (2015) 133–142.
[11] A. Alsger, B. Assemi, M. Mesbah, L. Ferreira, Validating and improving public transport origin–destination estimation algorithm using smart card fare data, Trans. Res. Part C: Emerg. Technologies 68 (2016) 490–506.
[12] N. Nassir, A. Khani, S.G. Lee, H. Noh, M. Hickman, Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system, Transp. Res. Rec. 2263 (2011) 140–150.
[13] F. Yan, C. Yang, S.V. Ukkusuri, Alighting stop determination using two-step algorithms in bus transit systems, Transportmetrica A Transport Sci. 15 (2019) 1522–1542.
[14] J.J. Barry, R. Freimer, H. Slavin, Use of entry-only automatic fare collection data to estimate linked transit trips in new york city, Transp. Res. Rec. 2112 (2009) 53–61.
[15] A.A. Alsger, M. Mesbah, L. Ferreira, H. Safi, Use of smart card fare data to estimate public transport origin–destination matrix, Transp. Res. Rec. 2535 (2015) 88–96.
[16] J. Zhao, A. Rahbee, N.H. Wilson, Estimating a rail passenger trip origin-destination matrix using automatic data collection systems, Comput. Aided Civ. Infrastruct. Eng. 22 (2007) 376–387.
[17] L. He, M. Trépanier, Estimating the destination of unlinked trips in transit smart card fare data, Transp. Res. Rec. 2535 (2015) 97–104.
[18] J.M. Farzin, Constructing an automated bus origin–destination matrix using farecard and global positioning system data in sao paulo, brazil, Transp. Res. Rec. 2072 (2008) 30–37.
[19] D. Li, Y. Lin, X. Zhao, H. Song, N. Zou, Estimating a transit passenger trip origin-destination matrix using automatic fare collection system, in: in: International Conference on Database Systems for Advanced Applications, Springer, 2021, pp. 502–513.
[20] J. Hora, T.G. Dias, A. Camanho, T. Sobral, Estimation of origin-destination matrices under automatic fare collection: the case study of porto transportation system, Transportation Research Procedia 27 (2017) 664–671.
[21] B. Assemi, A. Alsger, M. Moghaddam, M. Hickman, M. Mesbah, Improving alighting stop inference accuracy in the trip-chaining method using neural networks, Public Transport 12 (2020) 89–121.
[22] X. Liu, P. Van Hentenryck, X. Zhao, Optimization models for estimating transit network origin–destination flows with big transit data, J. Big Data Analytics in Transp. 3 (2021) 247–262.
[23] S. Lee, J. Lee, B. Bae, D. Nam, S. Cheon, Estimating destination of bus trips considering trip type characteristics, Appl. Sci. 11 (2021) 10415.
[24] J. Jung, K. Sohn, Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data, IET Intel. Transport Syst. 11 (2017) 334–339.
[25] D. Lei, X. Chen, L. Cheng, L. Zhang, P. Wang, K. Wang, Minimum entropy rate-improved trip-chain method for origin–destination estimation using smart card data, Trans. Res. Part C: Emerging Technologies 130 (2021) 103307.
[26] S. Cerqueira, E. Arsenio, R. Henriques, Is there any best practice principles to estimate bus alighting passengers from incomplete smart card transactions, in: Transport Research Arena Conference, 2022.
[27] S. Cerqueira, E. Arsenio, R. Henriques, Inference of dynamic origin–destination matrices with trip and transfer status from individual smart card data, Eur. Transport Res. Rev. 14 (2022) 1–18.

[28] A. Bhaskar, E. Chung, et al., Passenger segmentation using smart card data, IEEE Trans. Intell. Transp. Syst. 16 (2014) 1537–1548.
[29] K. Mohamed, E. Côme, L. Oukhellou, M. Verleysen, Clustering smart card data for urban mobility analysis, IEEE Trans. Intell. Transp. Syst. 18 (2016) 712–728.
[30] D. Luo, O. Cats, H. van Lint, Constructing transit origin–destination matrices with spatial clustering, Transp. Res. Rec. 2652 (2017) 39–49.