CIRRELT-2025-15

Improvement to Destination Estimation in AFC Systems

Melvin Huang
Martin Trépanier
Amaury Philippe

May 2025

Bureau de Montréal
Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec
Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

Improvement to Destination Estimation in AFC Systems

Melvin Huang*, **Martin Trépanier, Amaury Philippe**

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Polytechnique Montréal

Abstract. Smart card data in public transport provides a continuous flow of detailed information on passenger movements. Although it is often limited to entry transactions, this data can be enhanced by algorithms to estimate alighting locations and reconstruct origin-destination routes. Existing algorithms are generally effective at estimating destinations for most trips. However, some trips with missing information and not part of a sequence are more challenging. To address this, an improved destination estimation model is proposed combining criteria based on stop sequences and the historical transaction data of passengers.

Keywords: Origin and destination algorithm, Public transport systems, Smart card data, Passenger counting data.

Acknowledgements. The authors wish to acknowledge the funding from Keolis and MITACS, as well as Keolis Besançon team's collaboration for this study.

---

* Corresponding author: melvin.huang@polymtl.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2025

1. INTRODUCTION

The emergence of Smart Card Automated Fare Collection (SCAFC) systems in urban transport networks marks a significant milestone for the public transportation sector. Smart cards enable the continuous collection of individualized transactional data on public transport usage. These systems provide more flexible and secure ticketing while facilitating the gathering of vast quantities of transactional data rich in spatiotemporal information, such as time, location, and card type (Conklin and al. 2004). Smart card data is indispensable not only for revenue management but also for strategic, tactical, and operational planning in public transport networks (Pelletier and al., 2011). These data enable the evaluation of network performance through key indicators such as punctuality, service frequency, passenger load and passenger-kilometres (Trépanier and al., 2009). They also support passenger clustering (Morency and al., 2007) and trip purpose identification (Devillaine and al., 2012) aiming to enhance service efficiency and promote data-driven decision-making.

However, in systems that only capture entry validations ("tap-in"), only boarding point data is available, making it challenging to create complete origin-destination matrices essential for network planning (Kumar and al. 2018). To address this limitation, planners must use algorithms to estimate possible destinations based on travel sequences (Trépanier and al. 2007; Munizaga and Palma 2012). These algorithms reasonably assume that passengers alight at a stop near their next boarding stop. However, destinations not directly linked to travel sequences cannot be determined using this approach (Cheng and al. 2019). Extensions to existing algorithms aim to incorporate these unlinked trips, extracting more destinations and improving origin-destination matrices. A kernel density estimation approach is employed to evaluate the probabilities of destinations associated with these unidentifiable trips (He and Trépanier 2015). This extension increased the number of destinations found per transaction. For our algorithm, two new criteria based on stop sequences were integrated, along with two others inspired by the probabilistic method of He and Trépanier (2015) based on smart card history. The use of smart card data also extends to assessing transport network performance by measuring key indicators such as commercial speed and passenger-kilometres, offering planners insights to adjust transport supply to actual demand (Trépanier and al., 2009).

The article presents the proposed method and the data used in the case study. For the purpose of study, smart card transactions include those made with other means of payments as well (smart phone, tickets). The results of destination estimation, their validation using counting data. Finally, the article concludes with perspectives for future research.

2. METHODOLOGY

We begin by presenting the database used followed by a description of the classic Origin/Destination algorithm before introducing the improvements made to it.

2.1. Data source

This research was conducted in collaboration with Keolis, a major public transport operator in France. Besançon is a city of 120,000 inhabitants covering an area of 65.5 km². Keolis operates a fleet of buses and trams in this city. The smart card data used in this study corresponds to two weeks of transactions in June 2024 totaling over 750,000 transactions made by more than 75,000 cards. These transactions include not only validations performed with standard subscription cards but also those made with single-use tickets stored on smart cards as well as payments made by bank cards linked to a specific smart card identifier. Thus, all public transport users on the Keolis network are recorded in the AFC system through a unique smart card identifier except in cases of fare evasion (journeys made without validating a smart card). The data provided for this study includes transactions (card number, date, vehicle number, route number, direction, boarding stop and trip ID), passenger counts on both stations (boarding/alighting) and routes (on-board loads) and GTFS data of the service. The AFC system of the transport network demonstrates an overall satisfactory level of accuracy. However, some anomalies persist in the data. Approximately 8.39% of transactions lack information related to the trip identifier (trip ID). In addition, an estimated 0.60% of transactions contain an incorrect boarding stop identifier and 1.07% have an incorrect direction often due to drivers forgetting to update the direction information in the AFC system. To ensure the quality of the data used for analysis a preprocessing step is carried out. This process aims to associate each transaction with a valid trip ID from the GTFS data and to correct when necessary boarding stop identifiers and recorded directions.

2.2. Origin and destination algorithm

The proposed improvements build upon the classic algorithm developed by Trépanier and al. (2007). The indices used are as follows: *i* for the public transport user (unique smart card identification), *j* for the stop sequence number on a line, *r* for the user's route sequence number during the day and *k* for the day. A transport line *R* (in a single direction) is defined by an ordered set of stops *s*:

$$R = \{s^j\} \tag{1}$$

Thus, s_{rik}^{j} represents the *jth* stop of the *rth* route of user *i* on day *k*. Let us define the vanishing route, V_{rik}, as the sequence of stops where the user can alight after boarding a given route at stop *j* = *B*:

$$V_{rik} = \{s_{rik}^{j}\}, \forall j > B \tag{2}$$

Each s_{rik}^{j} corresponds to the event where user *i*, aboard the vehicle on route *r*, reaches stop *j* at time *t*. The trip J_{ik} of a public transport user can be modelled as a sequence of stop itineraries (vanishing routes) used during the day. Let N_k be the number of stop itineraries (vanishing routes) taken in a single day *k*.

$$J_{ik} = \{V_{rik}\},, r = 1 \dots N_k \tag{3}$$

Let $d(a, b)$ be the Euclidean distance between locations *a* and *b*. The model seeks to estimate the best location "the best location" for different values of *r* during the journey. The goal is to determine the destination d_{rik} in the stop itinerary in such a way that the distance to the next boarding stop for the next transaction, $s_{(r+1)ik}$ is minimized. We also define a maximum allowable distance *M* which is set to 1 kilometre.

$$\begin{gathered} d_{rik} = z \to \min_{z} d(s_{(r+1)ik}^{B}, z) \;\; z \in \{V_{rik}\}, \\ r < N_k \quad d(s_{(r+1)ik}^{B}, z) < M \end{gathered} \tag{4}$$

For the last trip of the day, we try to find a destination with the origin of the first trip of the day. If it does not work, we also try with the origin of the first trip of the next day. Figure 1 presents the classical algorithm.

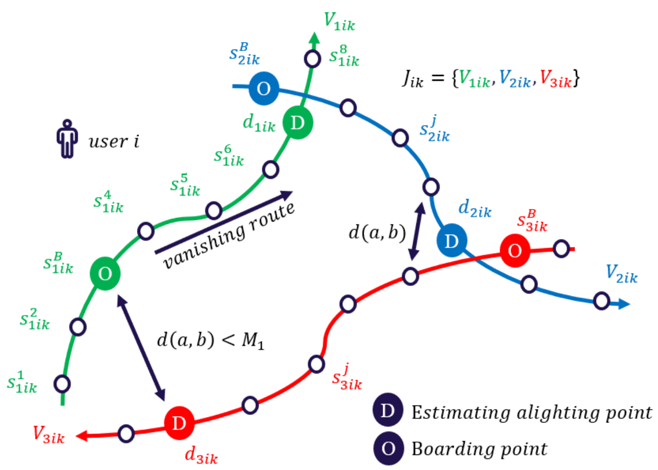

Figure 1. Classical Algorithm Model (inspired by Trépanier and al.).

2.3. Proposed improvements

All trips for which the destination cannot be estimated using equation (4) are classified as "other" trips (unlinked trips). New criteria are therefore proposed to more effectively estimate the destinations of these trips. The goal is to minimize the use of criteria based on the history of each user (a criterion that will be described later in the article), as information from transactions that are temporally close to the transactions under study seems more accurate than global history.

Previously, for the last transaction of the day assumptions were made to estimate the drop-off point: either the first stop of the same day was taken as the reference station or the first stop of the following day. If neither of these assumptions allows for a destination to be found, we now assume that the destination of the transaction is the stop closest to the boarding stop of the first transaction of the previous day provided the distance limit is respected.

Sometimes, the last transaction of a day also corresponds to the last transaction of a workweek. In this case, the next transaction may take place several days later for example after a weekend or another period of interruption. If the previous assumptions still do not allow for a destination to be estimated a new approach is considered: it is assumed that the destination of the last transaction can correspond to a stop located at a distance less than a predefined limit from the first transaction of a day between *k+2* and *k+7*. This assumption applies only if the transactions concern the same bus line but in the opposite direction.

Then, in the same way if the previous assumptions do not allow for a destination to be found it is assumed that the distance between the destination of the last transaction of the day and the last transaction of a day between *k-2* and *k-7* is less than a predefined distance limit with the condition that the transaction was made on the same bus line and in the opposite direction.

All trips for which the destination cannot be estimated using the previously seen equations will be subject to a new criterion called "historical" meaning they will be compared to the set of trips for the same user to identify a similar trip both spatially and temporally for which the destination would have been determined. The historical criterion developed is inspired by the work of He and Trépanier (2015) by proposing a new model to improve this criterion. To characterize the criterion, new indices are required: the type of day ($w\ et\ hw = \{0: weekday, 1: Weekend\}$, the time of day (p and hp = {Night (00:00 to 06:00), Morning Peak (06:00 to 09:00), Afternoon (09:00 to 15:00), Evening Peak (15:00 to 18:00), Evening (18:00 to 00:00)}), $H^B_{r,i,k,w,p}$ the set of potential historical destinations

$hd_x(r,i,K,w,p)$ for the map at the boarding stop B with $K = \{< day\ k\}$ and $y \in \{1,2,\dots,Y\}$ with Y: the number of unique potential historical destinations and N: the total number of potential historical destinations

$$H^B_{r,i,k,w,p} = \{hd_1(r,i,K,hw,hp),\dots,hd_X(r,i,K,hw,hp)\ if\ w = hw\ and\ p = hp\}$$
$$n_y = number\ of\ hd_y$$
$$P_y = \frac{n_y}{N}$$
$$d'''_{r,i,k} = z \to \max_z P_y\ \ z \in \{V_{r,i,k}\}\quad s^B_{r,i,k} = start(d'''_{r,i,k}) \tag{5}$$

The proposed model examines the history of the associated card to identify a similar transaction which is used to predict the potential destination of the current trip. Similarity is defined by the type of day and the time of day of the transaction. Similar transactions are then filtered to retain only those corresponding to the same line-direction. Finally, the most frequently observed destination among these transactions is assigned to the transaction under study. Figure 2 illustrates the historical criterion.

2.4. Weighted random sampling

Despite the improvements made to the algorithm, some trips cannot be processed even using the historical criterion. To overcome this limitation, the TAP (*tirage aléatoire pondéré*, Weighted Random Sampling) criterion has been integrated into the algorithm. This criterion utilizes the history of all available cards and acts as a last resort to handle trips whose destination remains undetermined.

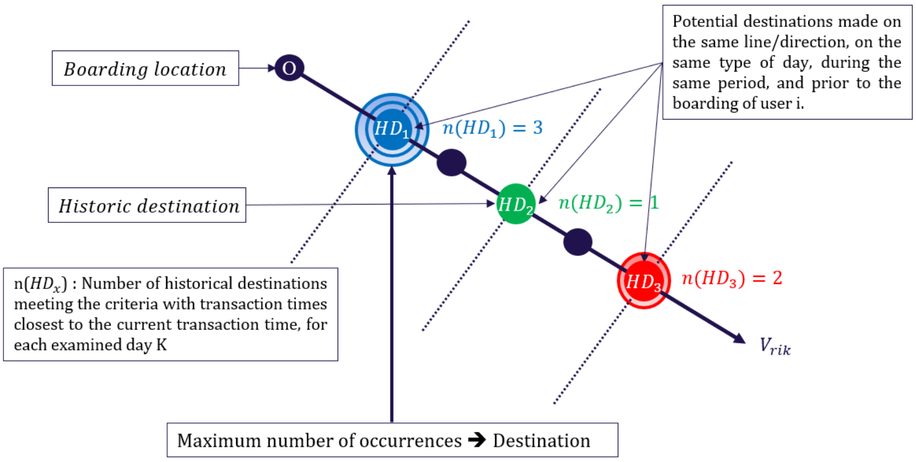

Figure 2. General description of the Historical criterion

The TAP criterion is inspired by the historical method but including all passengers: the list of potential historical destinations includes those where the transaction was made at the same time of day, the same type of day and on the same line-direction as the transaction under study. However, the difference lies in the final step of selecting the drop-off location among these potential destinations. A random number r uniformly distributed in the interval [0,1] is generated to make the final decision regarding the destination of the transaction under study. To facilitate the selection process a cumulative distribution C_n for each potential historical destination is defined:

$$C_y = \sum_{j=1}^{Y} P_j\ with\ C_Y = 1 \tag{6}$$

Thus, the historical destination hd_y is selected as the drop-off location of the transaction if r falls within the interval:

$$C_{y-1} < r \leq C_y \tag{7}$$

In other words, r is compared to the cumulative distribution to determine the "probability range" in which it falls. Figure 3 illustrates the TAP criterion.

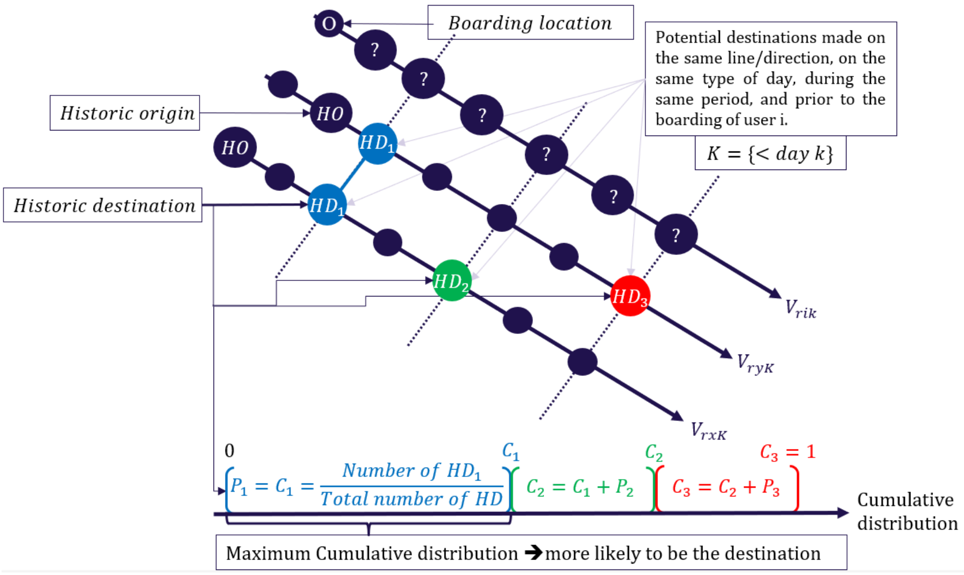

Figure 3. General description of the 'Weighted random selection' criterion

3. RESULTS

The Origin/Destination algorithm was applied to the Keolis database starting in June 2024 for a period of two weeks. In this analysis, each transaction was associated with a specific criterion, based on the following annotations:

- Criterion 1.1: the destination was identified during the trip sequence phase.
- Criteria 1.2 to 1.6: the destination was determined during the last trip of the day (also known as the return home, crit. 1.2), the first trip of the following day (crit. 1.3), the first trip of the previous day (crit. 1.4), the first trip of a day between j+2 and j+7 with the same line and opposite direction (crit. 1.5) and the last trip of a day between j-2 and j-7 with the same line and opposite direction (crit. 1.6).
- Criterion H (Historical): The destination was determined using the user's card history, with one or more potential destinations for the trip.
- Criterion TAP (Weighted Random Sampling): The destination was determined using the history of all cards in the database with one or more potential destinations.
- Criterion H|TAP: No destination could be identified.

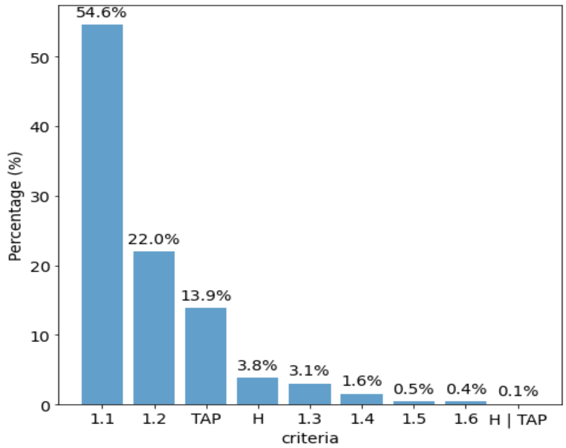

Figure 4. Percentage of occurrences of the different criteria for the Origin/Destination algorithm

3.1. Overall Result

Figure 4 shows the percentage of occurrences of the different criteria. The classical algorithm based solely on criteria 1.1, 1.2, 1.3 and the historical criterion can estimate the destinations of 83.5% of transactions. By integrating additional criteria 1.4, 1.5, 1.6 as well as the TAP criterion our improved algorithm can estimate approximately 16.4% more destinations thereby covering 99.4% of the trips that the initial algorithm was unable to resolve. Thus, only 0.1% of the trips remain unresolved. The main causes of non-resolution are an incorrect line or stop name in the data highlighted by the algorithm, an incorrect combination of the line and stop and the absence of historical data. It is worth noting that the percentage of use of the TAP criterion is relatively high compared to the H criterion. This can be explained by the similarities between the newly added criteria (1.4, 1.5 and 1.6) and the H criterion as they are based on a temporal characteristic.

3.2. Assessing the reliability of the origin and destination algorithm using counting data

To assess the reliability of the Origin/Destination estimation algorithm we used passenger count data at bus and tramway stations provided by Keolis covering the same period as the smart card transactions studied. To facilitate the comparison between the count data and the smart card validations an aggregation by trip ID was performed. An initial analysis revealed notable discrepancies between the two datasets mainly due to the presence of fare evasion. After cleaning the count data, only the tramway data were retained for comparison as they were considered more reliable than the bus data. However, even for trams certain anomalies persist such as positive or negative numbers of passengers recorded as remaining onboard at the end of a trip mainly due to detection errors by the counting system.

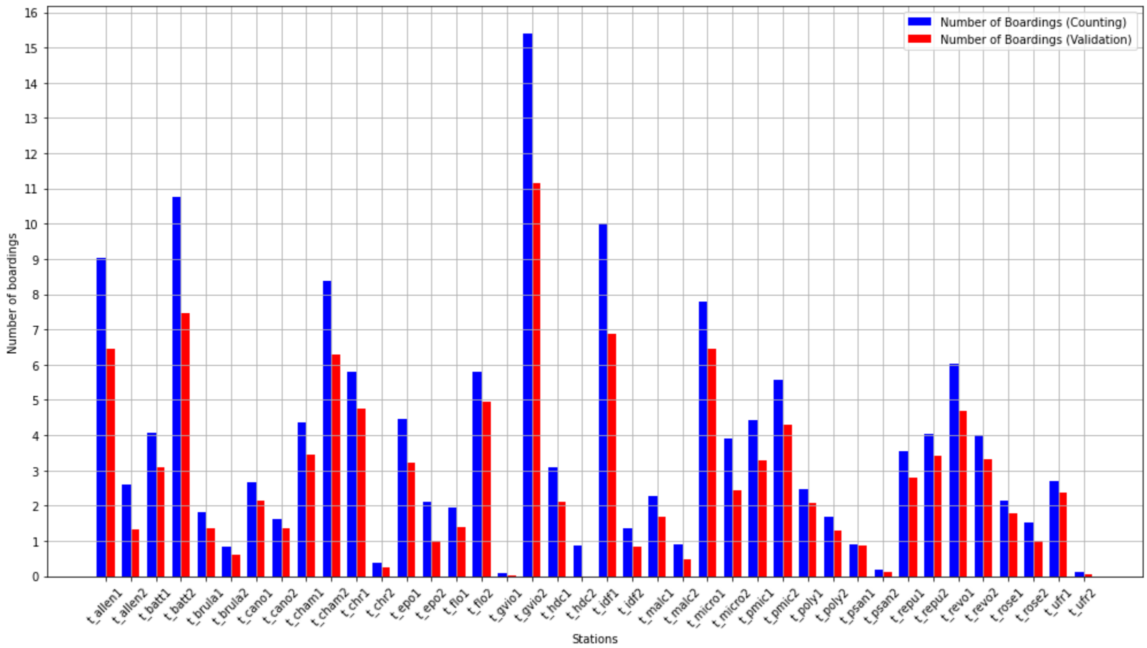

Figure 5. Comparison of average boardings from counting data and smart card validations, by Station for Line 2 of the Tram in Besançon, France

Figure 5 illustrates this comparison, each bar represents the average number of boardings per station (in blue for the count data and in red for the smart card validations). It appears that boardings recorded by the counting system are generally higher than those recorded by smart cards with a difference of 1 to 4 passengers depending on the station corresponding to approximately 20–25% variation. This discrepancy can be attributed to three main factors: fare evasion, a bias toward overcounting in the counting system and the presence of individuals exempted from validation (such as staff members or young children). However, comparisons based on averages may mask local variations where boardings recorded by the counting system are lower than smart card validations. Thus, for a more accurate observation it would be pertinent to compare boarding data on specific trips for instance by analyzing the same trip ID over several days.

3.3. Evaluating destination estimation accuracy using counting data

Similarly, to assess the reliability of destination estimation we compared the alighting estimates derived from smart card data with those from passenger count data from tram stations provided by Keolis over the same period on a station-by-station basis. If the average number of alightings from smart card data exceeds that from the count data it could indicate that the algorithm is not entirely reliable. In Figure 6 which illustrates this comparison (each bar represents the average number of alightings per station with blue for count data and red for smart card validations) we observe that out of 39 stations only 4 have a higher number of estimated alightings from smart card data compared to count data with a difference ranging from 1 to 2 passengers. While this suggests some limitations in the algorithm this conclusion must be tempered as the count data itself contains errors.

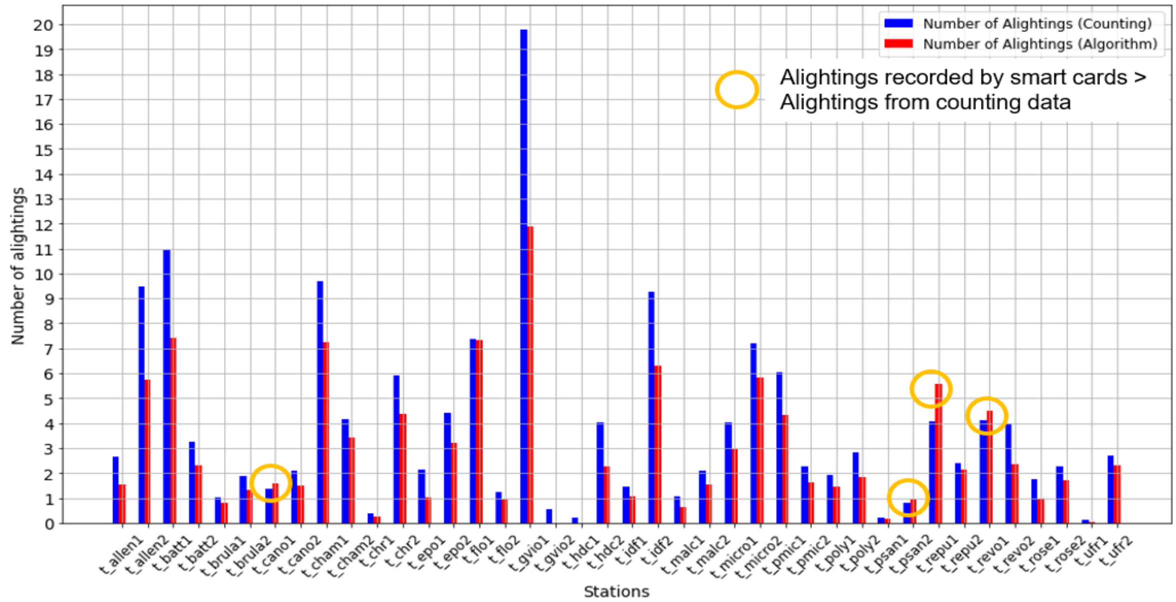

Figure 6. Comparison of average alightings from counting data and estimates by the Origin Destination Algorithm, by Station for Line 2 of the Tram in Besançon, France

For example, some stations report negative passengers counts due to system malfunctions. When averaging alightings per station these negative errors can distort the results leading to a lower average number of alightings for the count data compared to the smart card data. Additionally, 35 stations show a higher number of alightings in the count data than in the smart card data with differences ranging from 1 to 8 passengers. This result reinforces the apparent reliability of the algorithm but must also be interpreted with caution. Indeed, averaging results can introduce inaccuracies by masking specific variations in individual trips. It is also important to note that the algorithm's reliability primarily depends on its ability to accurately identify the destination of each individual transaction an aspect that cannot be fully assessed from Figure 6. To refine the evaluation, it would be more relevant to compare the destinations estimated from smart card data with the passenger count data for a specific trip identified by its trip ID. Furthermore, although the improved algorithm can identify the destination for 99.9% of transactions some uncertainty remains regarding the accuracy of these estimations due to the absence of tap-out data to validate them. However, comparing the results with passenger count data significantly reduces the risk of errors compared to an approach relying solely on smart card validations.

4. CONCLUSION

New criteria have been proposed to estimate the destinations of unlinked trips in a SCAFC system. Two of these criteria rely on stop sequences while the other two leverage the card's transaction history by assessing similarities with past transactions including a weighted random sampling for transactions that were unable to estimate using individual passenger assessment. Further investigations using alighting counting data are necessary to identify the criteria requiring adjustments or improvements. Currently, comparisons are based on average counting data but a more precise analysis focusing on specific trip_id would help determine which criteria contributed to identifying specific stations. This approach would also verify whether the surplus alightings are related to the historical or TAP criteria which include a probabilistic component. If so, these criteria could be adjusted to consider not only transaction similarity but also the volume of alightings at each station as recorded in the counting data. If the surplus alightings originate from other criteria recalibrating distance thresholds may be necessary as these thresholds can be influenced by variations in travel behaviours between cities. For the remaining 0.1% of trips without destinations (Criterion H|TAP) adjustments to the algorithm's conditions could help estimate the missing destinations. Possible solutions include increasing the distance limit and relaxing conditions related to temporal similarity such as the time of day or type of day although this may reduce accuracy.

ACKNOWLEDGMENTS

The authors wish to acknowledge the funding from Keolis and MITACS, as well as Keolis Besançon team's collaboration for this study.

REFERENCES

Cheng, Z., Trépanier, M. and Sun, L. (2019). Inferring Trip Destinations in Transit Smart Card Data Using a Probabilistic Topic Model. CIRRELT Report No. CIRRELT-2019-47, Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport, https://www.cirrelt.ca.

Conklin, J., Englisher, L. and Shammout, K. (2004) 'Transit Customer Response to Intelligent Transportation System Technologies: Survey of Northern Virginia Transit Riders', *Transportation Research Record: Journal of the Transportation Research Board*, 1887(1), pp. 172–182, https://doi.org/10.3141/1887-20.

Devillaine, F., Munizaga, M. and Trépanier, M. (2012) 'Detection of Activities of Public Transport Users by Analyzing Smart Card Data', *Transportation Research Record: Journal of the Transportation Research Board*, 2276(1), pp. 48–55, https://doi.org/10.3141/2276-06.

He, L. and Trépanier, M. (2015) 'Estimating the Destination of Unlinked Trips in Transit Smart Card Fare Data', *Transportation Research Record: Journal of the Transportation Research Board*, 2535(1), pp. 97–104, https://doi.org/10.3141/2535-11.

Kumar, P., Khani, A. and He, Q. (2018) 'A robust method for estimating transit passenger trajectories using automated data', *Transportation Research Part C: Emerging Technologies*, 95, pp. 731–747, https://doi.org/10.1016/j.trc.2018.08.006.

Morency, C., Trépanier, M. and Agard, B. (2007) 'Measuring transit use variability with smart-card data', *Transport Policy*, 14(3), pp. 193–203, https://doi.org/10.1016/j.tranpol.2007.01.001.

Munizaga, M.A. and Palma, C. (2012) 'Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile', *Transportation Research Part C: Emerging Technologies*, 24, pp. 9–18, https://doi.org/10.1016/j.trc.2012.01.007.

Pelletier, M.-P., Trépanier, M. and Morency, C. (2011) 'Smart card data use in public transit: A literature review', *Transportation Research Part C: Emerging Technologies*, 19(4), pp. 557–568, https://doi.org/10.1016/j.trc.2010.12.003.

Trépanier, M., Tranchant, N. and Chapleau, R. (2007) 'Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System', *Journal of Intelligent Transportation Systems*, 11(1), pp. 1–14, https://doi.org/10.1080/15472450601122256.

Trépanier, M., Morency, C. and Agard, B. (2009) 'Calculation of Transit Performance Measures Using Smartcard Data', *Journal of Public Transportation*, 12(1), pp. 79–96, https://doi.org/10.5038/2375-0901.12.1.5.