Onboard occupancy reconstruction in a public transport network through fraud modeling by combining AFC and APC data

Matteo Bénaud*[1], Baptiste Rérolle[1], and Martin Graive[1]

[1]Data Scientist @ CITiO, France

Short summary

Accurate estimation of ridership and fare evasion is essential for monitoring public transport networks. The combination of Automated Fare Collection and Automated Passenger Counting (APC) data delivers good results in fraud detection but its estimation on unequipped vehicles remains challenging. This paper proposes a new method to estimate fraudulent boardings and alightings on vehicles that are not equipped with APC, using transfer learning from equipped vehicles. Constrained linear regressions are fitted at multiple spatio-temporal granularities. To address data sparsity, a K-Nearest Neighbors approach transfers information across close spatio-temporal granularities. The method is evaluated on four public transport networks showing a significant improvement in onboard occupancy reconstruction compared to other existing approaches, and highlighting the importance of incorporating temporal dynamics in fraud estimation.
Keywords: Public transport, Fraud modeling, Linear regression, Automated Fare Collection, Automatic Passenger Counting, Transfer Learning

1 Introduction

An accurate estimate of the ridership is essential to understand how a public transport network is used. This enables transport operators and public authorities to adjust the level of service to meet demand. However, locating and quantifying fraudsters is a major challenge, as they represent a significant share of passengers while remaining largely undetected by AFC systems. As Barabino et al. (2020) shows in its review of fare evasion in public transport, the number of articles devoted to modeling and estimating fraud has increased in recent years, with several approaches now in use.

The traditional method to estimate fare evasion involves conducting an on-site survey, as described in Cantillo et al. (2022). Over the last few decades, new data sources have provided alternatives to these occasional and costly surveys. Nicodeme (2022) uses computer vision approaches to detect the posture of fraudsters entering a station or vehicle. Automated Fare Collection (AFC) data, as presented in Cevallos et al. (2021), can also be used for this purpose. A comparison with a fare evasion survey conducted in Lyon, France, demonstrates the effectiveness of the AFC-based approach and highlights the importance of combining AFC data with Automatic Passenger Counting (APC) data when available (see Egu & Bonnel (2020)). Indeed, the difference between the number of passengers boarding as recorded by the APC systems and the number recorded by the AFC provides an estimation of the number of fare evaders.

As APC systems are expensive, transportation networks are only equipping part of their vehicle fleets with them. Dib et al. (2023) has developed a method combining AFC and APC data to model fraud on vehicles not equipped with APC systems. A fraud rate is calculated and then applied to ticketing data at a station level. For stations where vehicles equipped with APC systems have passed through, the fraud rate is the average of the fraud rate measured on equipped vehicles. For stations where no vehicles equipped with APC systems have passed through, the fraud rate is computed from a geospatial regression based on stations with APC measures. This method has two main limitations. Firstly, Dib et al. (2023) fails to consider variations in fraud over time, calculating a uniform fraud rate for each station in the network. Secondly, since the calculated fraud rate is applied to the number of boarding validations, the model cannot account for fraudster boarding at stops where there are no ticketing validations. This paper offers a new method to estimate fraudulent boardings, alightings and therefore onboard occupancy in vehicles that are equipped with AFC systems but not APC systems, addressing the two main limitations of Dib et al. (2023) by including temporal variations in fraud and the possibility of only fraudster boarding at a stop.

2 Methodology

Data presentation

Automated Fare Collection AFC systems record all validation events across the network, that is, every time a passenger validates their ticket. The collected validations are linked to a vehicle, a stop and a datetime, and can be identified by a unique ticket or card identifier for each boarding.
However, the alighting information is usually unavailable in this data source. Therefore, AFC data must be processed to be converted into O-Ds or onboard occupancy.

Automatic Passenger Counts A subset of vehicles is equipped with APC systems, most often based on infrared sensors. These sensors provide measurements of boardings and alightings at each stop of the vehicle. Onboard occupancy at a stop is then calculated by subtracting the number of alightings from the number of boardings at the stops previously served by the course, where a course denotes a vehicle's trip along a line at a specific time.
Unlike AFC, APC captures all passengers, including fare evaders. However, APC data can be noisy or unbalanced, resulting in under or overestimation of onboard occupancy. It is the case for example in the situation where a different number of boardings and alightings has been recorded on one course. To address such inconsistencies, the process proposed by Cherrier et al. (2023), relying on linear optimization under constraints, is applied.

O-D inference from AFC and APC combination In this work, knowing the alighting stations is essential because it serves as the basis for calculating fraudulent alightings. To overcome the absence of alightings in AFC data, a four-step O-D reconstruction pipeline is implemented, based on both AFC and APC data:

- Trip chaining (inspired by Barry et al. (2002)): consecutive validations of the same user are linked to infer intermediate destinations. The underlying assumption is that the destination of a trip is close to the origin (validation) of the next one. This method can only be applied to trackable ticket types.
- Individual statistical inference (inspired by He & Trépanier (2015)): user-specific travel patterns are learned from historical data, allowing probabilistic inference of alighting stops.
- As described in Amblard et al. (2023), a Fisher's non-central hypergeometric sampling is used when APC data is available. Alighting stations are associated to boarding validations at a course level.
- When previous methods cannot be used, O-D matrices are built from the previously reconstituted trips and from the spatial distribution of validations not yet reconstituted, leading to a distribution of alightings for each entrance station. Alighting stations are then randomly sampled. This is done separately by weekday.

As this O-D reconstruction method is not a contribution of this paper, its results are not discussed in the following sections.
We estimate the number of fraudsters on vehicles equipped with APC devices by subtracting the reconstruction of boardings and alightings based on APC data, which includes all passengers, from that based on AFC data, which includes only passengers with a valid ticket.
This provides a proxy for the number of fraudsters at a course and stop levels, that is considered to be the ground truth.

Fraud inference model

The objective of the proposed method is to estimate fraud on vehicles not equipped with APC systems by leveraging information from equipped vehicles. The overall workflow is illustrated in Figure 1.
Traditional approaches typically rely on a fraud rate applied to validated boardings. However, such formulations implicitly assume that fraud can only occur when validations are observed. To relax this assumption, we instead model fraud using a linear regression framework that includes an intercept that allows for fraudulent boardings and alightings even without ticketing validations.

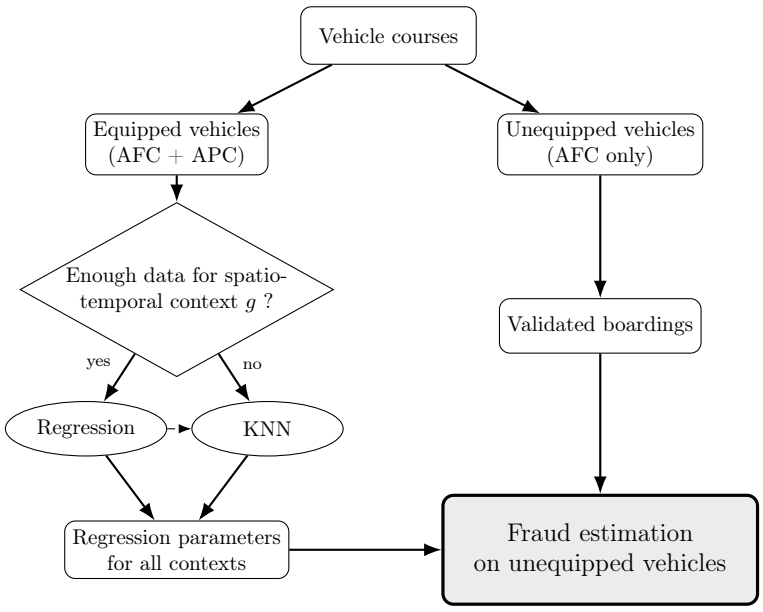


Figure 1: Workflow of the proposed algorithm. Regression parameters for each spatio-temporal context g are estimated from equipped vehicles and then transferred to unequipped vehicles to infer fraud and reconstruct fraudulent boardings and alightings.

In order to capture behavior variations at a low granularity, regression parameters are estimated for each context $g = (l, s, r, d, h)$ with l being the line, s the stop, r the direction, d the day, and h the hour.
Although AFC and APC data are available at a finer temporal resolution, observations are aggregated at the hourly level. This choice is motivated by a trade-off between temporal precision and statistical robustness: using hourly aggregation increases the number of observations per context, leading to more stable parameter estimation while still capturing meaningful temporal variations in fraud behavior. When sufficient data is available, a regression is trained. Otherwise, parameters are inferred from neighboring contexts using a K-Nearest Neighbors (KNN) approach.
For unequipped vehicles, fraud boardings are estimated using the inferred regression parameters. Fraud alightings are then inferred using the same modeling framework applied to reconstructed alighting flows. Vehicle onboard occupancy is subsequently reconstructed from these adjusted flows.

Regression model We model the number of fraudulent boardings at a given stop as a function of observed ticketing demand.
For each context g, we define:

$$F_g = \alpha_g + \beta_g B_g^V + \epsilon_g \tag{1}$$

where:

- F_g is the variable representing the number of fraudulent boardings,
- B_g^V is the variable representing the number of validated boardings,
- $\alpha_g \geq 0$ is the regression coefficient that captures the baseline fraud independent of validations,
- $\beta_g \geq 0$ is the regression coefficient that models validated boardings-dependent fraud,
- ϵ_g is a residual term, that the regression aims to minimize.

The positivity constraints ensure that the estimated number of fraudsters remains non-negative. This formulation extends classical fraud rate approaches by allowing non-zero fraud even when no validation occurs ($\alpha_g > 0$), thereby accounting for the fact that fraud can happen independently of the presence of fare-compliant users.

K-Nearest Neighbors extrapolation In many contexts, APC data is sparse or unavailable, either because some lines are not equipped with APC sensors or because too few equipped vehicles operate within a given context g.
To estimate parameters (α_g, β_g) in such cases, we rely on a K-Nearest Neighbors (KNN) approach. This choice is motivated by the assumption that similar spatio-temporal contexts tend to exhibit similar fraud behaviors. KNN provides a simple and effective way to exploit this structure by transferring information from neighboring contexts. The validity of this assumption is supported by the empirical performance observed in the section 3.
Each context is embedded into a joint spatio-temporal feature space:

$$x_g = (s_x, s_y, r_h h, r_w d) \tag{2}$$

where (s_x, s_y) are the spatial coordinates of the stop, h the hour of day, and d the weekday. The parameters r_h and r_w convert temporal differences into equivalent spatial distances, ensuring comparability across dimensions.
The regression parameters are estimated as:

$$(\alpha_g, \beta_g) = \frac{1}{K} \sum_{g' \in \mathcal{N}_K(g)} (\alpha_{g'}, \beta_{g'}) \tag{3}$$

where $\mathcal{N}_K(g)$ denotes the set of K nearest neighbors of g.
The KNN procedure relies on four hyperparameters:

- K, the number of neighboring contexts used for parameter estimation.
- m, the minimum number of observations required to fit a regression directly on context g. When fewer than m observations are available, KNN extrapolation is used.
- r_h, the scaling factor converting one hour difference into an equivalent spatial distance (in meters per hour), controlling the trade-off between temporal and spatial proximity.
- r_w, the scaling factor converting one weekday difference into an equivalent spatial distance (in meters per day), controlling the influence of weekly patterns.

These parameters jointly determine how information is transferred across contexts in the absence of local observations, and are tuned according to the method explained in the results section.

3 Results and discussion

Case study and experimental setup

Networks studied The study is conducted on three months of data, from April to June 2025, on four public transport networks. These networks were chosen because of their high coverage and quality of APC data, which require little treatment by the preprocessing algorithm presented above, thereby providing a solid ground truth. The agencies are :

- the Mistral network operated by RATP Dev Toulon Provence Méditerranée in the city of Toulon in southern France, with around 500 thousand people in the metropolitan area;
- Network A, a network in the Paris metropolitan area, with around 200 thousand people;
- Network B: a city in eastern France, with around 100 thousand people in the metropolitan area;
- Network C: a city in the Middle East, with around 5 million people in the metropolitan area.

These networks represent a wide range of operational contexts in terms of size, ridership, and APC coverage, allowing for robust evaluation of the proposed methodology.

Test dataset generation In order to compare our fraud estimation workflow to reference values, we restrict the analysis to courses equipped with APC data. In public transport fleets, some lines do not have any courses equipped with APC sensors, while other lines rely on partially equipped courses. To replicate this, we follow a two-step process to build the test dataset:

- randomly select a line, remove all related APC data, and repeat until at least 10% of the data is removed;
- randomly pick 15% of courses and remove the related APC counts.

This results in the following train-test splits (see Table 1).

Table 1: Train–test split proportions by network

City	Train	Test
Toulon	72.5%	27.5%
Network A	71.5%	28.5%
Network B	65.7%	34.3%
Network C	74.4%	25.6%

This allows us to evaluate our algorithm on the newly missing datapoints and compare it to the APC counts, considered as the ground truth.

Model evaluation The performance of the proposed method is evaluated using the weighted Mean Absolute Percentage Error (wMAPE), computed on the vehicle onboard occupancy at each stop. This metric is particularly relevant as ridership constitutes the primary quantity of interest for operational applications, such as vehicle sizing.
Compared to boarding-level metrics, onboard occupancy-based evaluation is less sensitive to temporal or spatial shifts in passenger assignment. For instance, assigning a passenger boarding to a nearby stop has a limited impact on onboard occupancy, whereas it would significantly affect boarding-based error metrics.
The wMAPE is defined as:

$$\text{wMAPE}(O, \tilde{O}) = \frac{\sum_{i=1}^{n} \left| O_i - \tilde{O}_i \right|}{\sum_{i=1}^{n} |O_i|} \tag{4}$$

where O_i and $\tilde{O}_i$ denote the observed and estimated onboard occupancy values, respectively.
The proposed method is compared against two benchmark approaches:

- A baseline method based on a uniform fraud rate applied across the entire network, computed as the global ratio between APC and AFC counts on the subset of equipped vehicles;
- The model proposed by Dib et al. (2023), which estimates station-level fraud rates and extrapolates missing values using geospatial regression.

Comparing the two methods described above allows us to assess the impact of the spatial component on fraud estimation. A comparison between our approach and Dib et al. (2023) method highlights the contribution of the temporal dimension in our reconstruction.

Results

Hyperparameter tuning As introduced before, the K-Nearest Neighbors extrapolation relies on four hyperparameters: K, m, r_h, and r_w. These parameters were selected through a grid search aimed at minimizing the discrepancy between inferred and observed fraud counts on the training dataset.
The explored ranges and corresponding optimal values are reported in Table 2.
The optimal number of nearest neighbors $K = 10$ indicates that estimating regression parameters from several neighboring contexts yields better results than relying on the single closest context only. This suggests that fraud behavior is not sufficiently stable at the finest granularity level and that moderate averaging across similar spatio-temporal contexts improves robustness.
The optimal value of the minimum number of observations required to fit a regression $m = 3$ seems low. This suggests that, even when only a small number of observations is available, fitting a regression directly on the target context remains preferable to extrapolating parameters exclusively

Table 2: Grid search intervals and optimal hyperparameter values for KNN extrapolation

Parameter	Minimum value	Maximum value	Optimal value
K	1	15	10
m	1	10	3
r_h	$0.5\,\mathrm{m.h^{-1}}$	$1000\,\mathrm{m.h^{-1}}$	$500\,\mathrm{m.h^{-1}}$
r_w	$0.5\,\mathrm{m.day^{-1}}$	$1000\,\mathrm{m.day^{-1}}$	$500\,\mathrm{m.day^{-1}}$

from neighboring contexts. In other words, sparse local information appears to be more informative than a purely transferred estimate.
Finally, the optimal scaling factor values, $r_h = 500m.h^{-1}$ and $r_w = 500m.d^{-1}$ mean that in the KNN feature space, a temporal difference of one hour or one weekday is treated as equivalent to a spatial distance of 500 meters. This result indicates that temporal proximity plays a role comparable to short-range geographical proximity when transferring regression parameters across contexts.

Global wMAPE results The onboard occupancy reconstruction results obtained on the four study networks are reported in Table 3.

Table 3: wMAPE of the different onboard occupancy reconstruction methods

Network	Baseline	Dib et al. (2023)	Proposed algorithm
Toulon	34.9%	33.7%	**31.6%**
Network A	38.9%	39.3%	**33.0%**
Network B	45.9%	38.2%	**36.4%**
Network C	81.1%	65.3%	**43.8%**
Average	50.2%	44.1%	**36.2%**

The results show that both Dib et al. (2023) and the proposed algorithm outperform the uniform baseline, with the proposed method consistently achieving the lowest error. On average, the wMAPE is reduced from 50.2% to 36.2%, indicating a substantial improvement in onboard occupancy reconstruction accuracy.
Compared with the results of Dib et al. (2023), the proposed method provides consistent gains across all networks.
The largest performance gaps are observed on Network C, where the wMAPE decreases from 81.1% on the baseline to 65.3% on Dib et al. (2023) method to 43.8% on the proposed algorithm. Due to its size, Network C encompasses a wide range of fraudulent behavior that more sophisticated methods are able to detect.

Temporal variations of fraud rates To evaluate the ability of the proposed method to reproduce variations in fraud behaviors over time, we analyze fraud rates over weekdays and hours of the day. Results are based on the busiest line of the Toulon network.
Figure 2 presents the evolution of fraud rates across weekdays on line 1 of the Toulon network. For confidentiality reasons, the exact labels on the fraud rate values are not displayed. The observed fraud rates (hatched bars) exhibit variations throughout the week, with a noticeable increase on Sundays.
As the baseline model (yellow bars) produces a constant fraud rate across all days, it fails to capture fraud rate variations over the week. The spatial model proposed by Dib et al. (2023) exhibits limited variability, which arises indirectly from differences in service patterns rather than from explicit temporal modeling. Consequently, it fails to capture the full amplitude of the observed temporal fluctuations.
In contrast, the proposed algorithm follows more closely the observed variations. On this particular line, the proposed algorithm is the only one to reflect Sunday's increase, reflecting its ability to explicitly account for temporal effects in the modeling process.
Figure 3 also illustrates fraud rate variations across hours of the day. A similar pattern is observed: while the baseline and spatial models exhibit little to no temporal dynamics, the proposed method

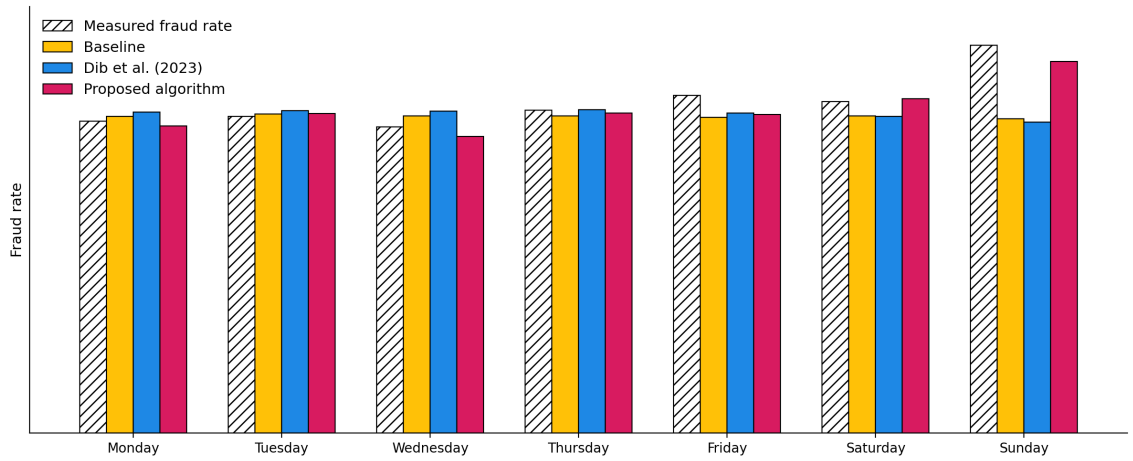


Figure 2: Evolution of fraud rates across weekdays on line 1 of the Toulon network. Fraud rate values on the y-axis have been removed for confidentiality reasons.

reproduces the observed hourly variations more accurately. In particular, it captures the increase in fraud rates during late afternoon and evening periods.

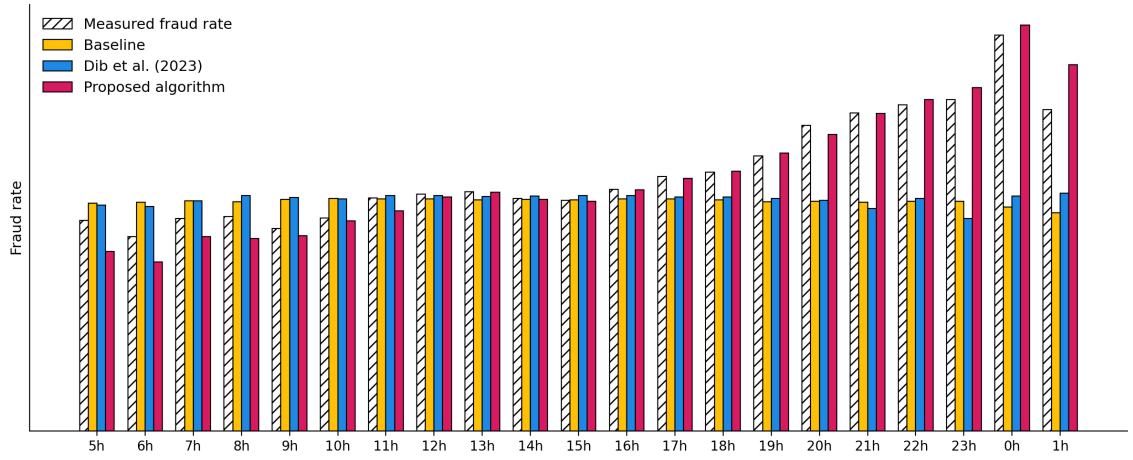


Figure 3: Evolution of fraud rates across hours on line 1 of the Toulon network. Fraud rate values on the y-axis have been removed for confidentiality reasons.

These results highlight the importance of incorporating temporal dimensions in fraud modeling. By accounting for both spatial and temporal variability, the proposed approach provides a more realistic representation of passenger behavior, which in turn leads to improved fraud estimation and more accurate onboard occupancy reconstruction.

Fraudulent boardings when no validations The Table 4 presents the percentage of fraudulent boardings inferred when no validations are observed at a vehicle stop. Both baseline and Dib et al. (2023) algorithms are always equal to 0.0%. As these two models directly apply a fraud rate to the number of validated boardings or alightings, it is expected that no fraudulent boardings are added when there are no validations. Our proposed method more closely follows the observed pattern, which is due to the use of a regression with a non-null intercept, instead of a simple rate.

4 Conclusions

This paper focuses on the problem of fraud estimation in public transport networks partially equipped with APC systems, through the combination of AFC and APC data.
Based on the assumption that fare evasion depends on the spatial and temporal context, the

Table 4: Ratio between the number of fraudulent boardings without any validated boardings at the vehicle stop over the total number of fraudulent boardings.

Network	True	Baseline	Dib et al. (2023)	Proposed algorithm
Toulon	28.7%	0.0%	0.0%	**25.4%**
Network A	20.8%	0.0%	0.0%	**23.9%**
Network B	24.5%	0.0%	0.0%	**29.7%**
Network C	47.2%	0.0%	0.0%	**52.4%**

proposed model splits ridership data by spatio-temporal groups to fit a linear regression between fare evasion and ticketing data from APC-equipped vehicles. When not enough data is available within a group, a KNN algorithm is used to transfer information across similar groups.
The proposed framework significantly improves the onboard occupancy reconstruction compared to a constant fraud-rate approach and Dib et al. (2023) method. The onboard occupancy wMAPE is reduced from 50.2% to 36.2% on average across four networks of different sizes and locations.
The proposed model succeeds in capturing fraud patterns observed by the APC systems and manages to add fraudulent boardings when no validated boardings are observed.

Acknowledgements

We want to thank the Mistral network of Toulon operated by RATP Dev Toulon Provence Méditerranée for allowing us to publish the results of our algorithms on their data. We also warmly thank the CITiO team for providing us with working and computing infrastructure, clean and formatted data, and domain expert knowledge to help us analyze results.

References

Amblard, V., Dib, A., Cherrier, N., & Barthe, G. (2023). A bayesian markov model for station-level origin-destination matrix reconstruction. In M.-R. Amini, S. Canu, A. Fischer, T. Guns, P. Kralj Novak, & G. Tsoumakas (Eds.), *Machine learning and knowledge discovery in databases* (pp. 538–553). Cham: Springer Nature Switzerland.

Barabino, B., Lai, C., & Olivo, A. (2020). Fare evasion in public transport systems: A review of the literature. *Public Transport*, *12*(1), 27–88.

Barry, J. J., Newhouser, R., Rahbee, A., & Sayeda, S. (2002). Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record*, *1817*(1), 183–187.

Cantillo, A., Raveau, S., & Muñoz, J. C. (2022). Fare evasion on public transport: Who, when, where and how? *Transportation Research Part A: Policy and Practice*, *156*, 285-295. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0965856421003141` doi: https://doi.org/10.1016/j.tra.2021.11.027

Cevallos, F., Turino, L., & Jin, X. (2021, May 01). *A synthesis on data mining methods and applications for automated fare collection (afc) data* (Final report, January 2020 - May 2021 ed.). Florida Department of Transportation. Retrieved from `https://rosap.ntl.bts.gov/view/dot/61846` (Tech Report)

Cherrier, N., Rérolle, B., Graive, M., Dib, A., & Schmitt, E. (2023). Context-aware automated passenger counting data denoising. In *2023 ieee 26th international conference on intelligent transportation systems (itsc)* (p. 1935-1942). doi: 10.1109/ITSC57777.2023.10422561

Dib, A., Cherrier, N., Graive, M., Rérolle, B., & Schmitt, E. (2023). Unified occupancy on a public transport network through combination of afc and apc data. In *2023 ieee 26th international conference on intelligent transportation systems (itsc)* (pp. 1963–1970).

Egu, O., & Bonnel, P. (2020). Can we estimate accurately fare evasion without a survey? results from a data comparison approach in lyon using fare collection data, fare inspection data and counting data. *Public Transport*, *12*(1), 1–26.

He, L., & Trépanier, M. (2015). Estimating the destination of unlinked trips in transit smart card fare data. *Transportation Research Record*, *2535*(1), 97–104.

Nicodeme, C. (2022). Fare-evasion detection at ticket gates using posture analysis. In *2022 ieee 25th international conference on intelligent transportation systems (itsc)* (pp. 1128–1132).