A Machine Learning Approach to adjust ridership computed from Wi-Fi data in Public Transport

Léa Fabre*[1], Caroline Bayart[2], Yacouba Kone[1], Ouassim Manout[1], and Patrick Bonnel[1]

[1]Urban Planning, Economics and Transport Laboratory (LAET), ENTPE, CNRS, University of Lyon, Vaulx-en-Velin, France
[2]Actuarial and Financial Sciences laboratory (LSAF), ISFA, University of Lyon, Lyon, France

Short summary

Wi-Fi data, collected from sensors placed on buses, seem promising for generating O-D matrices over a network. However, obtaining accurate bus ridership data is a challenge for public transport operators. Issues of completeness remain, as Wi-Fi sensors do not detect all signals emitted by connected objects in their vicinity, and some people do not own these devices. Data scaling is therefore a crucial step in the process of building O-D matrices from Wi-Fi data. In this work, four machine learning algorithms are compared to estimate the absolute values of passengers boarding and alighting at a bus stop based on Wi-Fi data and spatial and temporal characteristics. The results show that LGBM is the most relevant algorithm for generating accurate data.
Keywords: Correction, Deep learning, Machine learning, Passive data, Public Transit, Wi-Fi

1 Introduction

Understanding passenger behavior is a necessary prerequisite for planning and managing public transport (PT) systems (Wu et al., 2020). PT aims to provide users with safe, intelligent, reliable and efficient transportation services (Li et al., 2023; Haydari & Yılmaz, 2020). Mobility behavior has become less regular in recent years, due to profound socio-economic changes and more frequent disruptive events, such as the pandemic, (Škare et al., 2021), but also the emergence of new services and the uberization of processes. At the same time, transport studies have tended to incorporate new technologies and massive data.

Data can be collected from various sources like smart cards, telephony, Wi-Fi or Bluetooth sensors (Kaffash et al., 2021; Karami & Kashef, 2020), mostly passively, thanks to the development of new connected objects such as smartphones, smartwatches, connected headphones and so on. Wi-Fi data first appeared in mobility studies in the years 2010s, and the literature shows that they are a promising way to capture mobility behaviors (Dunlap et al., 2016; Blogg et al., 2010). They are collected using electromagnetic wave sensors, that detect connected objects in their environment, as they search for an access point to connect to (Pu et al., 2021). The connected objects are not personally identified, and the MAC addresses are often pseudonymized for confidentiality reasons (CNIL, 2021). In PT, sensors placed in vehicles and coupled with a GPS module can provide information about the origin and destination of passengers' trips. Being able to link the origin and the destination of a passenger trip, Wi-Fi sensors can be used to build dynamic O-D matrices. From this point of view, they have the potential to outperform traditional automatic passenger counting systems in understanding mobility patterns (Nitti et al., 2020). Wi-Fi sensors are used in several mobility studies, especially in PT, either to represent bus loads, or the number of passengers boarding or alighting at a stop (Paradeda et al., 2019; Hidayat et al., 2020) or to estimate O-D matrices (Pu et al., 2021; Nitti et al., 2020). Most of these studies have focused on the process of sorting Wi-Fi signals, to separate those belonging to real bus passengers from parasitic signals and to improve the representation of mobility behaviors (Jalali, 2019; Fabre et al., 2023).

However, the use of Wi-Fi data raises several challenges in terms of completeness: some passengers do not have a connected object or have many, some devices are not detected by bus sensors, and it is rare for all the buses in a network to be equipped with Wi-Fi sensors. Nitti et al. (2020) highlight that the percentage of ownership of a connected object varies according to social position and age, two key determinants of mobility behaviors. This finding suggests that there are differences in the

ownership of connected objects, depending on the time of day and the bus stop considered. This bias may lead to an under or overestimation of the number of people present in the vehicles if no weighting correction is applied, as shown in Kurkcu & Ozbay (2017). In addition, Franssens (2010) noted that some interference phenomena can affect signal detection, which is more likely to occur during peak hours and in city centers.
To reflect the mobility behavior of PT users, it seems important to take spatial and temporal variables into account when scaling Wi-Fi data. This paper aims to meet this objective by identifying and comparing different machine learning algorithms to weight Wi-Fi data. Machine learning and artificial intelligence algorithms make it possible to introduce a wide range of model inputs, including variables related to the spatio-temporal context. For example, Pu et al. (2021) use Random Forest regression to estimate PT ridership based on time variables. The methodology provides estimates that are fairly close to the observed data, but has only been applied to a small sample. The work of Chang et al. (2023) focuses on estimating online ridership and develops a model that combines various machine learning algorithms. The results show that the model underperforms in certain areas where people are expected to carry fewer connected objects.
This work compares and ranks different methods to weight PT ridership estimated from Wi-Fi data. Wi-Fi data correction is an essential step in the construction of reliable and representative O-D matrices. As O-D surveys are not conducted regularly enough, we use optical counts as reference data to scale boarding and alighting passenger numbers. Subsequently, the O-D matrix structure derived thanks to previously introduced methodologies Fabre et al. (2023) can be applied to the weighted set of trips starting or ending at each stop. The paper proposes a reliable method for estimating the actual number of boarding and alighting counts from Wi-Fi data. Its originality lies in the use of machine learning to have a weighting process that depends not only on the collected signals, but also on their spatio-temporal characteristics.

2 Methodology

The Wi-Fi data used in this work is first filtered to distinguish the signals emitted by real bus passengers from interfering signals, which are removed from the database. This step follows the methodology described in Fabre et al. (2023).
The general architecture of the method is shown in Figure 1. Optical counts are used as the "real" value. They provide the number of people getting on and off at each stop. The goal is to match the number of Wi-Fi detected passengers at a stop with the number of people at that stop as determined by optical counts. Two models are implemented, one for boarding passengers and one for alighting passengers.

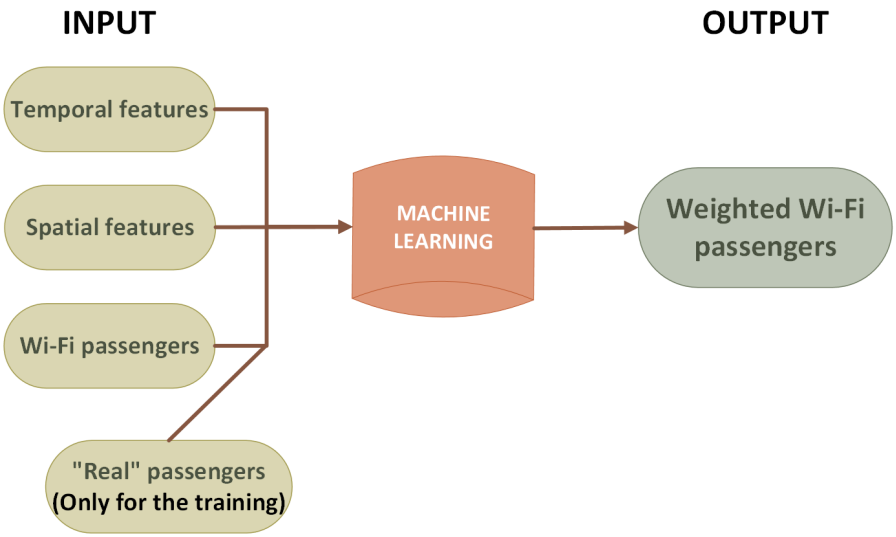

Figure 1: *Input and output variables for weighting Wi-Fi data*

We consider the number of people getting on and off at each stop detected by the Wi-Fi sensors to be the new MAC addresses and the last MAC addresses detected at each stop, respectively. The new MAC addresses at each stop correspond to the number of MAC addresses detected at

the stop that were not previously detected, and the last MAC addresses at each stop correspond to the number of MAC addresses detected at the stop that were not subsequently detected (i.e., the last time this object is seen at a stop). For this, we developed a method to assign signals to the nearest stop and to determine whether each of these signals belongs to a previously (or lastly) detected object or not. These variables are assumed to be a proxy for the real number of boarding and alighting passengers at each stop.
Exploratory analysis of the Wi-Fi data shows that the ratio between observed and detected MAC addresses varies between stops and according to hour, week and month of the year (Figure 2). We, therefore, include different spatio-temporal variables to account for this variability in the weighting process. Temporal variables are derived from the date and time of the observations and include day of the week, month of the year, peak hour information, etc. Spatial variables are derived from additional data sources and are merged thanks to the coordinates of each bus stop. They include the number of connections available at each stop and the population living around the stop (national de la statistique et des études économiques, 2016). In order to select the input variables for the models, some feature engineering was done for the prediction of boarding passengers and for the prediction of alighting passengers. A correlation analysis was then performed to avoid including correlated features in the models. The final selected variables are gathered in Table 1

Table 1: Selected features

Variable	**Type**	**Definition**
$nb_{new\ mac}$	Int	Number of new MACs detected
$nb_{last\ mac}$	Int	Number of last MACs detected
day	Int	Day of the year from 1 to 365
$weekday$	Int	Day of the week from 1 to 7
c_{ind}	Float	Number of persons living within a 200m square of the stop
$corres_{semirapid}$	Int	Number of possible connections with a BRT
$corres_{street}$	Int	Number of possible connections with a bus

Several algorithms are then used to predict the number of people getting on and off a bus. The goal is to compare the models and select the one that best replicates observations from optical counts. Linear Regression (LR) is used as a basis for the comparison. Three other models are run: Random Forest (RF), Light Gradient Boosting Machine (LGBM) and Multi-Layer Perceptron (MLP) (Hastie et al., 2009). All of them are compatible with high-dimensional data, allow for feature evaluation, and usually give very good predictions while maintaining fast execution. Grid search is used to find the optimal hyperparameters for RF and LGBM, trial and error was used to set the MLP hyperparameters, and no hyperparameters are needed for the LR. For the sake of brevity, these hyperparameters are not detailed in this abstract.

3 Results and discussion

Case study

This study used *Laflowbox*, a Wi-Fi data collection sensor developed by Explain, a French transport planning consultancy firm. The data were collected in Rouen, France, the largest city in the Rouen Normandie conurbation (70 municipalities, 494,000 inhabitants). Wi-Fi sensors were installed in buses belonging to the Transport Est Ouest Rouennais (TEOR) network, the main network in Rouen, which consists of four lines. In 2022, six different buses traveling on this network were each equipped with a sensor that collected data continuously throughout the year.
By calculating the daily number of passengers boarding and alighting, it is possible to observe differences in the relationship between Wi-Fi data and optical count data, both from one period of the year to another and from one stop to another. Figure 2 shows the daily boarding passengers throughout the year 2022 with optical counts and Wi-Fi data for two lines of the network. It appears that Wi-Fi data and optical counts are not linearly related over the year, nor from one line to another. This supports the need to find a method for weighting Wi-Fi data that takes into account spatial and temporal variability.

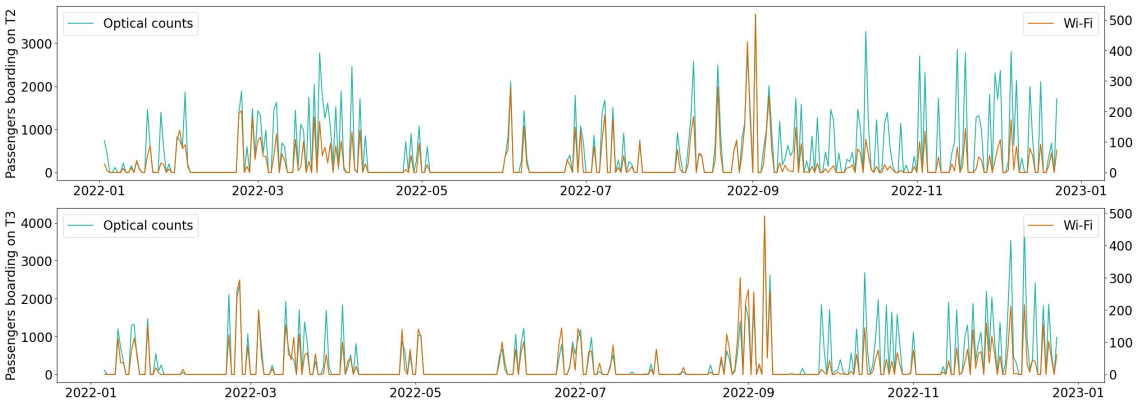

Figure 2: *Boarding passengers on T2 and T3 lines with Wi-Fi and optical counting*

Results

As discussed in the methodology, four models are considered for predicting boarding and alighting passengers at a stop from Wi-Fi data (LR, RF, LGBM and MLP). Three performance indices are computed to evaluate the quality of the proposed models. The coefficient of determination (R^2), the Root Mean Square Error ($RMSE$) and the Mean Absolute Error (MAE). The performance is evaluated on a daily basis, i.e. replication of daily boarding and alighting passengers number. To do this, we have added all the boarding (resp. alighting) passengers observed for the trips made on the same day.
The results for the four selected models are summarized in Table 2. The LGBM algorithm is found to outperform the three other algorithms in both cases (boarding and alighting), followed by the RF, NN and lastly LR. With LGBM $RMSE = 16.51$ for the prediction of boarding passengers and $RMSE = 17.29$ for the prediction of alighting passengers. Also, the maximums predicted with LGBM are close to those obtained with ground truth.

Table 2: Metrics for the different models implemented to predict boarding and alighting passengers (day scale)

	Model	R^2	$RMSE$	MAE	min	max
Boarding passengers	**LR**	0.61	37.98	20.96	1.0	520.0
	RF	0.89	20.39	12.50	0.0	800.0
	LGBM	0.93	16.51	10.04	0.0	832.0
	NN	0.79	27.65	16.55	0.0	902.0
Alighting passengers	**LR**	0.54	43.66	23.03	1.0	507.0
	RF	0.91	19.11	11.70	0.0	875.0
	LGBM	0.93	17.29	10.50	0.0	873.0
	NN	0.76	31.66	17.72	0.0	960.0

Based on the previous results, the Light Gradient Boosting Machine algorithm was selected to predict passengers boarding and alighting at each stop using Wi-Fi data. The following figures provide more insight into the predictions obtained with LGBM. The analysis is presented here for the prediction of boarding passengers only. Figure 3 (a) shows the predicted number of boarding passengers versus the actual number of boarding passengers for each observation in the database. The red line is the linear regression of the predicted passengers as a function of the observed passengers, and the black dashed line is the curve of the equation $y = x$. Figure 3 (b) shows the distribution of absolute errors between the predicted and observed number of boarding passengers. 96% of the observations are predicted with an absolute error between -30 and 30 passengers per day, which is very low. These two figures highlight the very good matching between observed and predicted boarding passengers.
The SHAP values (Hastie et al., 2009) show that the model relies mostly on the number of available

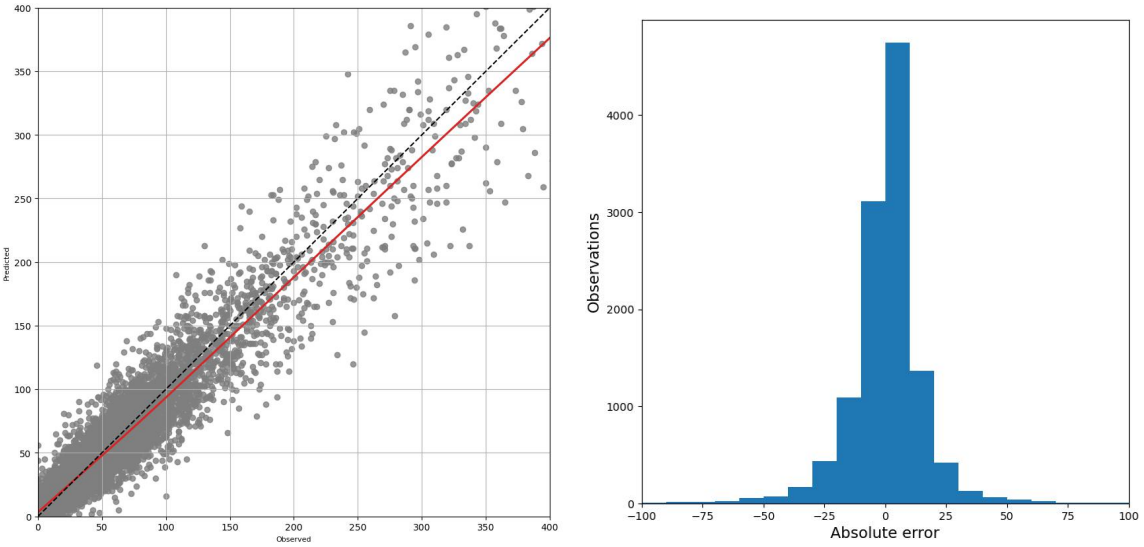

Figure 3: *(a) Linear regression of predicted with LGBM versus observed boarding passengers: (b) Distribution of absolute errors made on the daily predictions of boarding passengers with LGBM.*

connections ($corres_{street}$), the number of residents around the stop, and then equally on the number of new MAC addresses detected and the day of the year (features with the greatest influence). The impact of each of these features on the predictions is shown in Figure 4 (a), (b), (c) and (d). The number of predicted boarding passengers increases as the number of new MAC addresses detected increases until it reaches the value of four new MAC addresses detected at a stop for a single bus journey. It also increases significantly when the stop offers more than 4 connections. Concerning the temporal variables, the number of predicted boarding passengers is higher in March and in the fall. The decrease in August is also significant. The number of inhabitants also has an impact on the results, although the trend is less pronounced. There is a slight decrease in the number of predicted boarding passengers for the stops with 200 to 300 or very few inhabitants nearby.

4 Conclusions

This research deals with the use of Wi-Fi sensors in PT. While most studies on this topic focus on the filtering of noisy signals, there is another important step to take in order to derive accurate O-D matrices from Wi-Fi data, both in relative and absolute terms. This is the weighting of the data, required to obtain the correct passenger volumes. To address this issue, this paper aims to develop a methodology to weight Wi-Fi data in a way that takes into account the spatial and temporal variability in the scaling process. Four machine learning algorithms are compared : the Linear Regression, the Random Forest, the Light Gradient Boosting Machine and the Multi-Layer Perceptron. The models take as input the Wi-Fi detections at each bus stop, as well as spatial and temporal variables. The models are trained using optical counts as ground truth data. The experiment was conducted on a one-year long data collection in several buses of the TEOR network (Rouen Normandie Metropole, France) equipped with Wi-Fi sensors. The present study has achieved its objective, as several algorithms give encouraging results. The algorithm that leads to the best predictions of boarding and alighting passengers is the LGBM with lower $RMSE$ and higher R^2 between predicted and observed values. Thanks to this method, it is possible to weight the number of passengers boarding and alighting at a bus stop, taking into account the geographical location of the bus stop and the temporality. This can be done continuously with Wi-Fi detections only, since the model is already trained and performs well on different datasets (unless there is a very strong disruption in the network). Further work could include applying the weighted volumes to the previously obtained Wi-Fi O-D matrix structure (Fabre et al., 2023) to be able to continuously study trip volumes and structure, simply with Wi-Fi data.

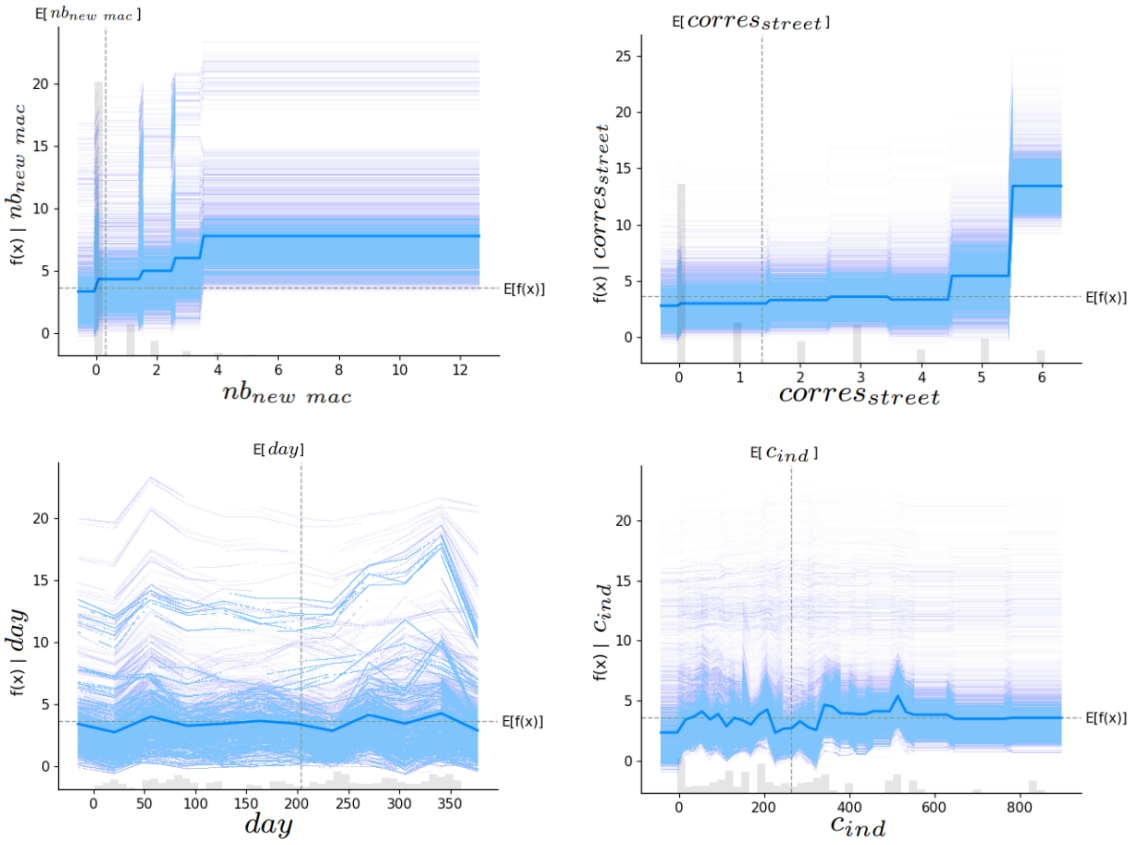

Figure 4: *Partial dependence plot for predicting boarding passengers with LGBM (a) Feature $nb_{new\ mac}$; (b) Feature $corres_{street}$; (c) Feature day; (d) Feature c_{ind}.*

Acknowledgements

The authors acknowledge EXPLAIN, the LAET and the LSAF for the financial support. The authors also thank EXPLAIN and the Métropole Rouen Normandie for providing data.
This proposition is part of a work in the process of submission to a journal.

References

Blogg, M., Semler, C., Hingorani, M., & Troutbeck, R. (2010). Travel time and origin-destination data collection using bluetooth mac address readers. In *Australasian transport research forum* (Vol. 36, p. 15).

Chang, W., Huang, B., Jia, B., Li, W., & Xu, G. (2023). Online public transit ridership monitoring through passive wifi sensing. *IEEE Transactions on Intelligent Transportation Systems*, *24*(7), 7025-7034. doi: 10.1109/TITS.2023.3257835

CNIL. (2021). *La loi Informatique et Libertés.*

Dunlap, M., Li, Z., Henrickson, K., & Wang, Y. (2016). Estimation of origin and destination information from Bluetooth and Wi-Fi sensing for transit. *Transportation Research Record*, *2595*(1), 11–17.

Fabre, L., Bayart, C., Bonnel, P., & Mony, N. (2023). The potential of wi-fi data to estimate bus passenger mobility. *Technological Forecasting and Social Change*, *192*, 122509. doi: https://doi.org/10.1016/j.techfore.2023.122509

Franssens, A. (2010). *Impact of multiple inquires on the bluetooth discovery process : And its application to localization* (Info:Eu-Repo/Semantics/masterThesis). University of Twente.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Haydari, A., & Yılmaz, Y. (2020). Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*(1), 11–32.

Hidayat, A., Terabe, S., & Yaginuma, H. (2020). Estimating bus passenger volume based on a wi-fi scanner survey. *Transportation Research Interdisciplinary Perspectives*, *6*, 100142. doi: https://doi.org/10.1016/j.trip.2020.100142

Jalali, S. (2019). *Estimating Bus Passengers' Origin-Destination of Travel Route Using Data Analytics on Wi-Fi and Bluetooth Signals* (Unpublished doctoral dissertation). Université d'Ottawa/University of Ottawa.

Kaffash, S., Nguyen, A. T., & Zhu, J. (2021). Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. *International Journal of Production Economics*, *231*, 107868. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0925527320302279` doi: https://doi.org/10.1016/j.ijpe.2020.107868

Karami, Z., & Kashef, R. (2020). Smart transportation planning: Data, models, and algorithms. *Transportation Engineering*, *2*, 100013. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2666691X20300142` doi: https://doi.org/10.1016/j.treng.2020.100013

Kurkcu, A., & Ozbay, K. (2017). Estimating pedestrian densities, wait times, and flows with wi-fi and bluetooth sensors. *Transportation Research Record*, *2644*(1), 72–82. (tex.eprint: https://doi.org/10.3141/2644-09)

Li, B., Guo, T., Li, R., Wang, Y., Gandomi, A. H., & Chen, F. (2023). Self-adaptive predictive passenger flow modelling for large-scale railway systems. *IEEE Internet of Things Journal*.

national de la statistique et des études économiques, I. (2016). *Données carroyées à 200 mètres.*

Nitti, M., Pinna, F., Pintor, L., Pilloni, V., & Barabino, B. (2020, January). iABACUS: A Wi-Fi-Based Automatic Bus Passenger Counting System. *Energies*, *13*(6), 1446. Retrieved 2020-08-25, from `https://www.mdpi.com/1996-1073/13/6/1446` (Number: 6 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/en13061446

Paradeda, D. B., Junior, W. K., & Carlson, R. C. (2019, November). Bus passenger counts using Wi-Fi signals: some cautionary findings. *TRANSPORTES*, *27*(3), 115–130. Retrieved 2020-09-21, from `https://revistatransportes.org.br/anpet/article/view/2039` (Number: 3) doi: 10.14295/transportes.v27i3.2039

Pu, Z., Zhu, M., Li, W., Cui, Z., Guo, X., & Wang, Y. (2021, January). Monitoring public transit ridership flow by passively sensing wi-fi and bluetooth mobile devices. *IEEE Internet of Things Journal*, *8*(1), 474–486.

Škare, M., Soriano, D. R., & Porada-Rochoń, M. (2021). Impact of covid-19 on the travel and tourism industry. *Technological Forecasting and Social Change*, *163*, 120469.

Wu, W., Xia, Y., & Jin, W. (2020). Predicting bus passenger flow and prioritizing influential factors using multi-source data: Scaled stacking gradient boosting decision trees. *IEEE Transactions on Intelligent Transportation Systems*, *22*(4), 2510–2523.