The Potential of Wi-Fi Data to Estimate Bus Passenger Mobility

Léa Fabre, Caroline Bayart, Patrick Bonnel, Nicolas Mony

LABORATOIRE AMÉNAGEMENT ÉCONOMIE TRANSPORTS
TRANSPORT URBAN PLANNING ECONOMICS LABORATORY

WORKING PAPERS DU LAET

NUMÉRO
2022/02

The Potential of Wi-Fi Data to Estimate Bus Passenger Mobility

Léa FABRE
Caroline BAYART
Patrick BONNEL
Nicolas MONY

Using technologies such as Wi-Fi and Bluetooth allows to gather passive mobility data, useful for ensuring the sustainable development of transport infrastructures. The challenge of passive data collection is to be able to identify relevant data. Our research presents interesting solutions for sorting the transmitted signals and reconstructing quality Origin-Destination matrices. Its originality consists not only in comparing the results with those of other data sources, but also in proposing a methodology that can be reproduced. Thanks to a partitioning algorithm, it is possible to automatically distinguish passengers from non-passengers to get transit ridership flow and O-D matrices. The findings show that this algorithm provides concrete and replicable solutions to transport operators for understanding travel demand.

Keywords: Travel behavior, passive tracking, data clustering, Wi-Fi/Bluetooth sensors, trajectory reconstruction, mobile devices data, data quality

LABORATOIRE AMÉNAGEMENT ÉCONOMIE TRANSPORTS
TRANSPORT URBAN PLANNING ECONOMICS LABORATORY

NUMÉRO
2022/02

The Potential of Wi-Fi Data to Estimate Bus Passenger Mobility

Léa FABRE
Explain, F-69004, Lyon, France - Laboratoire de Sciences Actuarielle et Financière - Univ Lyon, ENTPE, LAET, F-69120, VAULX-EN-VELIN, France

Caroline BAYART
Laboratoire de Sciences Actuarielle et Financière, ISFA, Université Claude Bernard Lyon 1, Lyon, France

Patrick BONNEL
Univ Lyon, ENTPE, LAET, F-69120, VAULX-EN-VELIN, France

Nicolas MONY
Explain, F-69004, Lyon, France

Juillet 2022

ISSN : 2741-8103

Laboratoire Aménagement Économie Transports
MSH Lyon St-Etienne
14, Avenue Berthelot
F-69363 Lyon Cedex 07 France

The Potential of Wi-Fi Data to Estimate Bus Passenger Mobility

Fabre Léa (Corresponding author)
Explain - Laboratoire de Sciences Actuarielle et Financière - Laboratoire Aménagement Economie Transport
Université Lumière Lyon 2, Lyon, France, 69000
Email: lea.fabre@entpe.fr

Caroline Bayart
Laboratoire de Sciences Actuarielle et Financière, ISFA
Université Claude Bernard Lyon 1, Lyon, France, 69000
Email: caroline.bayard@univ-lyon1.fr

Patrick Bonnel
Laboratoire Aménagement Economie Transport, ENTPE
Université Lumière Lyon 2, Lyon, France, 69000
Email: patrick.bonnel@entpe.fr

Nicolas Mony
Explain
Lyon, France, 69004
Email: nmony@explainconsultancy.com

Declarations of interest: none

Submission date : 20/01/2022

Abstract

Last decades have been marked by deep socio-economic transformations, an uneven evolution of main modal shares in urban areas and new transport modes emerging. All these changes have a strong and direct impact on individual mobility behaviors. In this context, using technologies such as Wi-Fi and Bluetooth allows to gather passive mobility data, useful for ensuring the sustainable development of transport infrastructures. The challenge of passive data collection is to be able to identify relevant data. Our research presents interesting solutions for sorting the transmitted signals and reconstructing quality Origin-Destination matrices. Its originality consists not only in comparing the results with those of other data sources, but also in proposing a methodology that can be reproduced. An experiment was conducted in the metropole of Rouen, a medium-sized French city with a dense urban bus network. Some buses have been equipped with "Laflowbox", a French electromagnetic wave sensor able to anonymously capture the MAC addresses of passengers' portable devices. Thanks to a partitioning algorithm, it is possible to automatically distinguish passengers from non-passengers to get transit ridership flow and O-D matrices. The findings show that this algorithm is more efficient than existing filtering methods. They provide concrete and replicable solutions to transport operators for understanding travel demand and managing the quality of service.

Keywords: Travel behavior, passive tracking, data clustering, Wi-Fi/Bluetooth sensors, trajectory reconstruction, mobile devices data, data quality

1. Introduction

Last decades have been marked by several socio-economic transformations (demographic growth, population ageing, urbanization…). In addition, we observe an uneven evolution of the modal share of private car and public transports in urban areas, as well as new transport modes emerging (car sharing, e-bikes, electric scooters...) and recently, events such as the pandemic. Several studies (Axhausen et al., 2002; Borkowski et al., 2021; Deschaintres, 2018; Eisenmann et al., 2021) showed that these changes have a strong and direct impact on mobility behaviors. In this context, it is more necessary than ever to obtain accurate and representative data in order to feed complex transport demand models and ensure a smarter and environmentally friendly mobility (Bonnel, 2004).

The use of technologies such as Wi-Fi and Bluetooth (Hidayat et al., 2018; Ji et al., 2017; Nitti et al., 2020; Pu et al., 2021) is known as the most recent way to capture data in transport planning (Nitti et al., 2020). It has been democratized thanks to the spread of portable electronic devices (smartphones, headphones, laptops, smartwatch…), as nowadays, more than 80% of the population possesses a Wi-Fi or Bluetooth connected object (Pu et al., 2021). Wi-Fi and Bluetooth data are collected from sensors that gather numerous data on a regular basis, which leads to a finer temporal characterization of mobility behaviors. Compared to traditional surveys, data gathering and processing are faster, less expensive, and easy to implement as they don't need any action of individuals. So, Wi-Fi sensors appear as a promising opportunity to collect mobility data. However, the big quantity of data offered by passive collection does not necessarily means a more accurate illustration of mobility behaviors. Some work needs to be done to analyze representativeness of Wi-Fi and Bluetooth data. This technology presents some limits: some passengers do not hold any connected devices, and these are not always detected by the sensors present in bus. In addition, it remains challenging to identify the signals coming from bus passengers among the huge number of detected signals. In most past experiences, threshold values are defined to address this issue, which do not allow for spatial variability and limit replication. Lastly, most research using WI-FI sensors to capture urban mobility data are restricted to a small field of survey, on a short period. Hence, much work remains to be done to obtain a method for passively collecting quality mobility data in urban public transport.

The paper aims to address these limits by using algorithms to automatically identify the Wi-Fi signals emitted by connected objects from bus passengers, build tracks and generate Origin-Destination matrices. Data are collected from "Laflowbox", an electromagnetic wave sensor developed by Explain (French consulting firm in transport planning) to measure individual mobility in public transport. "Laflowbox" sensors are embedded in buses and can track users equipped with connected devices, without identifying them personally. The study was conducted during one week in October 2018, on a whole bus line of the city of Rouen, a medium-sized French city which takes advantage of a dense bus network. The first main objective of this research consists in identifying methods to select signals coming from bus passengers which are no longer dependent on threshold values, specific to each experiment. We intend to develop a method which could be replicable on other bus lines and other agglomerations. The second main objective is to increase the scale of the study by confronting the results with several sources of real data commonly used in transport planning (optical counts, O-D onboard survey and smartcard data) on an entire bus line, on a whole day of operation. This will also confirm the quality of mobility data gathered continuously in public transport by Wi-Fi sensors and validate the use of clustering algorithms to get O-D matrices from these passive data.

The remainder of this paper is organized as follows. Section 2 summarizes the state of the art on the use of Wi-Fi data to construct O-D matrices and raises literature gaps. Section 3 presents the Wi-Fi sensing device, named "Laflowbox", and the data processing. The methods used to get O-D matrices from gathered data are presented in section 4. Section 5 highlights the numerical results of the clustering algorithm used to select bus passengers signals as well as a comparison with several sources of real data. These results are then discussed in section 6, before concluding the study and proposing future research issues.

2. Related work

2.1. The evolution of data sources to describe mobility

The Household Travel Surveys (HTS), conducted at regular intervals in large and medium-sized cities, as well as National Transport and Travel Surveys are essential for understanding daily mobility behaviors. They draw a portrait of all the trips made by a representative sample of inhabitants of a predefined perimeter. Performed in face to face or by phone, these surveys are extremely costly for local authorities and only record trips declared by the respondent over a relatively short period (one or two days), every ten years on average. In order to compensate drawbacks of the national ones, other surveys are often performed, for public transport these are mainly origin-destination surveys made on board.

Last decades, survey data collection methods have evolved with the aim to get reliable and accurate enough information. A goal of these new data collection methods is also to get a sample closer to the surveyed population, to solve the representativeness issue encountered with traditional survey modes (Bonnel et al., 2015; Bonnel and Munizaga, 2018). Hence, data survey methods have evolved to get a better understanding of mobility behaviors variability, but also to incorporate new technologies such as web, GPS devices and smartphones (Montini et al., 2015; Patterson and Fitzsimmons, 2016). These new media remain periodic and still have to face high costs, non-answer or processing challenges. Also, while several survey modes are available, the problem of data comparability has been the subject of much research (Bayart and Bonnel, 2012).

Since a few years, democratization of passive data collection is observed in transport planning. This includes ticketing, telephony data and even some artificial intelligence processes. They provide a very large amount of continuous data, but do not give any details about trips and their attributes (Egu and Bonnel, 2020). It therefore seems appropriate to look at new technologies that enable transport operators to gather reliable and massive mobility data at a lower cost.

2.2. The potential of Wi-Fi sensors to gather mobility data

Wi-Fi or Bluetooth sensors have emerged in transport planning literature since the years 2010s and seems to be a promising way to capture mobility (Blogg et al., 2010; Dunlap et al., 2016; Ji et al., 2017; Malinovskiy et al., 2012). In these studies, Wi-Fi sensors detect the unique Media Access Control (MAC) addresses of connected objects if their Wi-Fi function is turned on (Pu et al., 2021). This way of detecting active Wi-Fi interfaces is easy to implement because only one sensor by bus is needed and no specific calibration is required (Michau et al., 2013). Generally, data collected by Wi-Fi sensors come with a unique MAC address, the precise time of detection, and the signal strength (Fukuda et al., 2017; Ji et al., 2017). When a MAC address is detected several times at different positions, an origin and a destination stop can be affected thanks to a pairing with a GPS. Lastly, it is possible to build Origin-Destination matrices with a high temporal and spatial granularity (Calabrese et al., 2013; Hidayat et al., 2018). By getting boarding and alighting stops besides the count of people onboard, these sensors bring a more complete picture of individual mobility (Nitti et al., 2020). Compared to traditional surveys, Wi-Fi and Bluetooth sensors are easy to implement, as they allow real-time entry of data, without any physical effort (Kyritsis, 2017). Indeed, this way of detecting connected objects does not need any action of the user and is totally passive (Nitti et al., 2020). Data gathering and data processing are thus faster. Lastly, this solution remains less expensive, especially for large amounts of data (Traunmueller et al., 2017). Despite the cost associated with the purchase of sensors, they can be used over a long period with a marginal cost per record. Thus, considering the democratization of data processing and internet access, Wi-Fi and Bluetooth sensors are promised with a fast development. However, this technology presents some limits, which should be considered. First, the sensors do not capture strictly all the objects present in the bus, as their detection range can be weakened by several parameters. In their work, Michau et al. (2013) have highlighted difficulties linked to bad weather conditions as well as obstacles or too many connected objects inside the vehicles. The higher the number of detectable objects around the sensor, the more inference phenomena

can affect signal detection (Franssens, 2010). A sensor which is not well positioned in the bus can lead to non-complete signal detection. For that reason, some studies use several sensors in the same bus (Mehmood et al., 2019). Two other constraints must be considered by analysts: some passengers do not hold any connected objects, while others have several (Nitti et al., 2020). This may lead to an under-estimation or an overestimation of the number of individuals present in the bus and bias the counts. This misrepresentation can be emphasized depending on the category of people (elder people are less often in possession of a connected object). The penetration rate of connected objects is also varying a lot depending on the country and more generally depending on economic growth. In underdeveloped countries, under-estimation of bus passengers is going to be observed, whereas in developed countries, the misrepresentation is coming from duplicate possession of connected objects (Nitti et al., 2020). Finally, the main challenge of Wi-Fi sensors is to identify the signals coming from bus passengers among the huge number of detected signals. The detection range of Wi-Fi sensors is larger than the vehicles, and many connected objects not belonging to passengers but to people near the bus can be detected. For example, pedestrians waiting at the bus stop or on the sidewalk, as well as cyclists passing. Vehicles on the road are also likely to have cell phones or connected watches whose Wi-Fi signal will be detected by sensors. The number of parasite signal is generally huge so that the final share of signals used to build Origin-Destination matrices is very low. Thus, it seems interesting to have a methodology to filter signals which does not depend on the operator.

2.3. The classification of Wi-Fi signals

Several studies (Araghi et al., 2012; Dunlap et al., 2016; Ji et al., 2017; Kurkcu and Ozbay, 2017; Myrvoll et al., 2017; Nitti et al., 2020) overcome this problem using threshold values to eliminate "unwanted signals" (minimal number/duration of the detection, signal strength…). If this method makes it possible to obtain quality data, corresponding to real bus passenger trips, threshold values are hard to determine and require an in-depth study of the territory beforehand (Pu et al., 2021). Some studies are not so optimistic about the capacity of threshold values for this step. Fukuda et al. (2017) showed that with this method, some objects not belonging to users of the bus were included in the O-D matrix. It is also not straightforward to choose the variables to determine if a signal belong to a passenger or not. Whereas signal strength and duration of the detection are most of the time preferred (Ji et al., 2017; Oransirikul et al., 2019), for Nitti et al. (2020), speed would be a better indicator to determine if the object is inside the bus or not. Also, use of threshold values to select signals is more subject to ignore temporal variability. As an example, if signal strength is decreasing when the crowd is heavy, the threshold value should not be the same at peak hour and during the rest of the day (Namaki Araghi et al., 2015). To the knowledge of the authors, only two recent studies refer to alternative methods to separate bus passengers from non-passengers. Afshari et al. (2019) used hierarchical clustering and K-means algorithm while Pu et al. (2021) chose fuzzy C means algorithm. Following a comparative analysis of the trips recorded by the sensor with real data, the authors mentioned the good quality of filtered signals. However, in both studies, the ground truth data consist of manual counts by an interviewer who identifies passengers using a hand-held sensor, which is switched on when the bus leaves and switched off at its terminus. As a result, these studies only consider data collected on a single bus and on only a few trips.

2.4. The lack of complete network representation

Most of the studies using Wi-Fi data to build Origin-Destination matrices focus on a single bus line or on a small territory. However, it is shown by Mishalani et al. (2016) that O-D matrices constructed with Wi-Fi data are closer to the ones constructed with ground truth data when the comparison is made on an aggregation of several journeys. Dunlap et al. (2016) limit their study to a line operating on a university campus. For Ji et al. (2017), the experiment was hold during several days but only a short period (between 8 and 9 am, on weekdays only) and only one bus following the same route was equipped. Moreover, it is mentioned in this work that in these conditions, individual mobility can be well estimated by Wi-Fi sensors only when it is merged with ticketing data. Also, the two previously mentioned studies, which did not use

threshold values to separate the signals, placed their Wi-Fi sensor in a single bus. Afshari et al. (2019) captured data on a single route but 6 times (during six days, from 4 pm to 5:30 pm). Pu et al. (2021) chose to survey three routes, three times each. In order to gather ground truth data for comparison, they used a surveyor counting passengers in the bus. All these experiments do not allow for a complete representation of mobility on the network, neither for temporal variations.

From this literature review emerged a need to analyze the quality of mobility data gathered in public transport by Wi-Fi sensors. First the data need to be processed automatically to be replicable to other field experiments. Second, the experiment should not be limited to a single bus over a short period of time, to be more representative and to allow comparison of the data with those commonly used in transportation planning.

3. From field to Wi-Fi dataset

We present the scope of the study and the sensing device proposed, before detailing the steps to process the data.

3.1. Data collection and study area

The data collection sensor used in this study is called "Laflowbox". It measures mobility thanks to an electromagnetic wave sensor which detects connected objects (smartphone, smartwatch…). Several versions of "Laflowbox" have been used since its development in 2015. They differ from each other by the technology used to detect objects (Wi-Fi or Bluetooth), their antenna coverage, their size, their autonomy.... The device used for this research is a small case directly plugged to the bus, so that it is powered as soon as the bus engine starts. The box also contains a small antenna to detect connected objects and a GPS module that will allow, in the end, to infer an origin and a destination to these objects and to build O-D matrices.

Every connected object emits Wi-Fi signals regularly, in order to be aware of the surrounding access points. The frequency of these emissions depends on the brand and the type of connected objects but also on the activity of the latest. Indeed, a smartphone which is being used (applications, internet…) will emits more frequently than a smartphone at rest in user's pocket. The emission frequency is not constant. "Laflowbox" detects these kinds of signals, without interacting with the emitting objects. The sensor is invisible from the connected objects, which makes the data collection completely passive. In parallel to this signal's detection, "Laflowbox" lists its GPS coordinates every second thanks to the GPS module. A box produces two kinds of data: data related to the detected connected objects and GPS data. The Wi-Fi file is made of an anonymized MAC address, a timestamp, and the signal strength for each observation. The GPS file is made of a timestamp, a latitude, and a longitude for each observation. A new line is created in the Wi-Fi file every time a connected object is detected by the sensor and, similarly, a GPS observation is appended to the GPS file every second. GPS data and Wi-Fi data can be linked thanks to a common timestamp.

Mac addresses are considered as personal data by the General Data Protection Regulation (GDPR) (Article 4(1)). However, it is also mentioned in Article 5.1.(b) that personal data could be collected for "specified, explicit and legitimate purposes including scientific or historical research purposes or statistical purposes" as long as it fulfills appropriate conditions (General Data Protection Regulation, n.d.). Hence MAC addresses collection with "Laflowbox" follows the CNIL requirements (CNIL, n.d.) as the MAC addresses are pseudonymized in very short delays. Users of the bus are also aware of the presence of this device through posters.

Data were gathered in Rouen, a large city in the Normandy region (North of France). With a bit more than 111 000 inhabitants Rouen is the biggest city of the metropole Rouen Normandie (70 cities, 494 000 inhabitants) and a rather important national economic hub, mainly thanks to its seaport, the fifth of France. "Laflowbox" sensors have been placed in buses of the TEOR (Transport Est Ouest Rouennais) network, the main network of Rouen with the subway line (**figure 1**). It is composed of 4 lines (but at the

time where the survey was conducted, the fourth line was not yet functioning), covering 65 stops and serving 7 other cities. Three of the four lines of the network share a common section of 10 stops in the city center of Rouen. Bus frequency is between 3 and 8 min during peak hour and less than 2 min on the common section. The first line, T1, holds 15 stops and is about 8 km long. The second line, T2, is much longer with 30 stops for over 30 km. Finally, the third line holds 27 stops and is about 14 km long. T1 entirely circulates on dedicated lanes, whereas this type of lane represents only one third of the T2 and T3 lines. All of this makes the TEOR network very attractive. Indeed, with more than 56 000 trips a day, the TEOR gathers 30% of the traffic flow and observes a constant increase since its implementation.

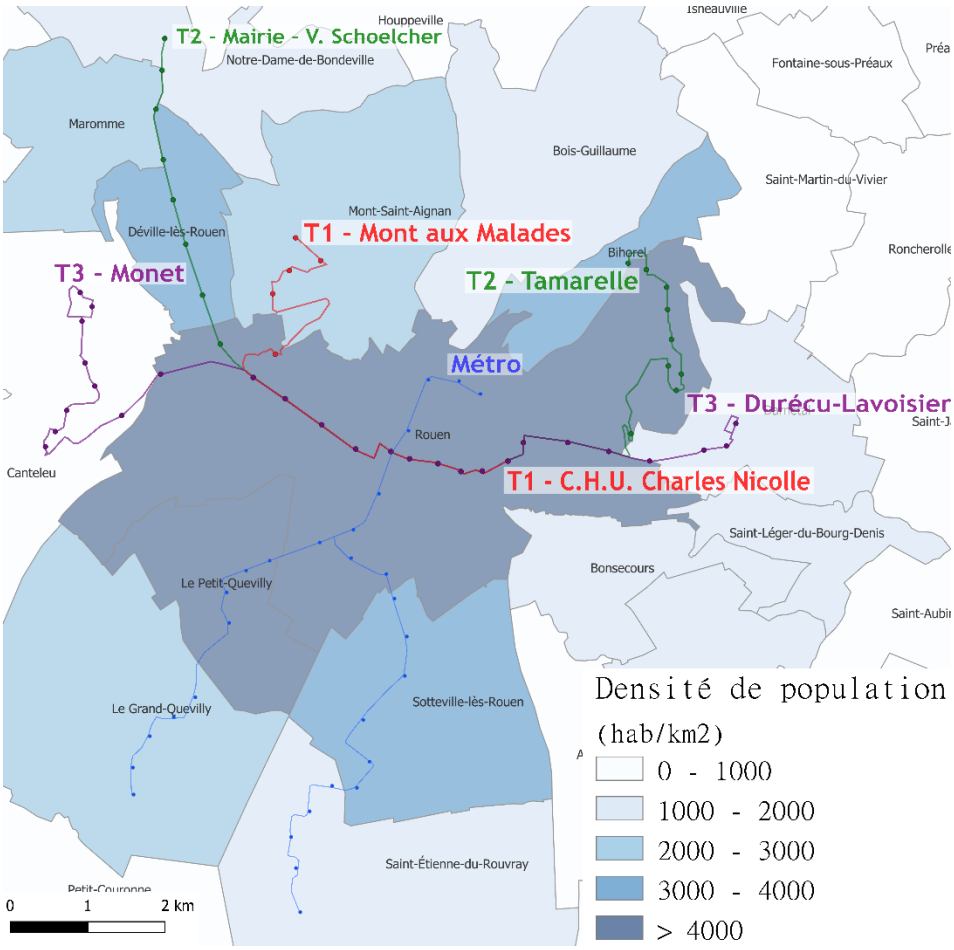


Figure 1: Main network of Rouen

3.2. Dataset and pre-processing

In October 2018, 16 sensors were placed in 16 different buses circulating on this network. They collected data continuously from Wednesday the 10th to Tuesday the 16th of October, representing over 12 million signals. The research is based on data collected on Thursday, October 11, 2018. This choice is motivated by the availability of other survey data on that day, allowing the authors to compare "Laflowbox" results with ground truth data. In this context, a database of 2 409 942 signals has been recorded. Then, some pre-processing has been carried out to clean up the data file.

First, some MAC addresses are removed from the database because they are identified as nonmoving objects (i.e. shops around the trajectory of the bus…). Then, by merging the Wi-Fi and the GPS files, a latitude and a longitude are associated with each observation in the Wi-Fi file. The next step is to create a new entity called a "track" that is the succession of Wi-Fi signals detected by a box, emitted by the

same object, and separated one from each other by less than 600 seconds. The limit was initially set to a longer time (the time needed to run through the longest line of the network) but too many tracks corresponded to several O-Ds. Then, ten minutes appeared to be reasonably long, even for objects with a very low emission frequency, and at the same time short enough not to build a track with several O-Ds. Some variables are computed both for the objects and the tracks, and other variables are computed for each signal (**table 1**). These variables have been selected after looking at the literature and the variable used when observing Wi-Fi data. The last pre-processing step consists in removing some values considered as "extreme" for any network. Thus, tracks with null distance or number of detections less than or equal to 3 are removed because it is not relevant to build a path with 3 points or less. Either they are all grouped, and the origin is the same as the destination, either they are very spaced and it is very likely that the object doesn't follow the bus path. Objects for which the average signal strength is positive are also removed (Wi-Fi signal strengths range from -30 to -90 dBm). Finally, we get rid of objects for which the track length is greater than 7200 seconds, 2 hours being much too long for a trip on this network.

Table 1: Variables computed

Variable	Object	Track	Signal
Total number of detections	X	X	
Total duration	X	X	
Distance travelled		X	
Time between two consecutive signals			X
Distance between two consecutive signals			X
Distance from barycenter of the track			X
Detection frequency	X	X	
Sum, standard deviation, variance, mean, median and interquartile range of the distance from the barycenter of the track		X	
Standard deviation, variance, mean, median and interquartile range of the signal strength		X	
Standard deviation, variance, mean, median and interquartile range of the time lapse between two consecutive signals		X	
Standard deviation, variance, mean, median and interquartile range of the distance between two consecutive signals		X	

After these pre-processing steps, the database finally contains 510 583 signals, 51 476 tracks and 37 371 objects. **Table 2** describes the number of operable observations recorded by each box on the 11th of October 2018. The number of signals detected can vary a lot from one box to another with 12 323 signals for box 45 and 47 007 for box 60. The number of operable signals depends on the proportion of parasitic signals detected but also on the time spent at the bus depot. It is likely that the bus equipped with box 53 did not travel a lot on the reference day.

Table 2: Wi-Fi detections by box

Box	Number of signals	Number of tracks	Number of objects	Box	Number of signals	Number of tracks	Number of objects
11	19 126	1 840	1 556	**45**	12 323	1 293	1 076
16	43 170	4 351	2 975	**50**	37 147	3 572	2 653
19	34 414	3 317	2 445	**53**	14	2	2
20	40 180	4 245	2 922	**54**	31 203	3 023	2 354
25	36 135	3 710	2 720	**60**	47 007	5 002	3 340
26	27 808	2 725	2 078	**76**	45 017	4 543	3 117
27	21 510	2 127	1 809	**87**	34 169	3 543	2 520
31	40 196	4 043	2 842	**99**	41 164	4 140	2 962

In this paper, Wi-Fi data are compared with reliable behavioral data, to validate the results obtained with "Laflowbox". Several datasets are available for the T3 line on the 11th of October 2018, hence the comparison will focus on line T3. First, a face-to-face Origin-Destination on-board survey. During this survey, the interviewer asked the passengers about their trip: mode, purpose at destination, departure time, boarding and alighting stops and any connections. Profile data such as the age and gender of the respondent are also provided. Finally, 12 626 observations have been recorded in both directions of the Line T3. This survey is supplemented with two other data sources: optical counts and ticketing. Optical counts are coming from cameras positioned in some buses, which count the number of people boarding and alighting at each stop. This device does not link a boarding to an alighting for a dedicated passenger, but provides, for the entire line, the number of boardings and the number of alightings at each bus stop. On the reference day, one bus was equipped with both this system and a Wi-Fi sensor "Laflowbox". Ticketing data concerns the validations made with smartcards when passengers board the bus. This dataset provides the number of validations for each stop and each trip on the reference day.

4. Methods

The methods used to select the Wi-Fi signals in order to build O-D matrices and to compare them to ground-truth data are explained in the following section.

4.2. Signals classification

As a first step, signals coming from actual bus passengers and unwanted signals are differentiated. This step consists in eliminating interfering signals (people waiting at a bus stop but not boarding in the vehicle, people cycling on the road, people in their car following the bus, and so on). It also aims at removing some signals coming from static connected objects, such as a computer in a shop or an office and close enough to the bus to be detected by the sensor.

To select signals emitted by bus passengers professionals usually use, threshold values for key variables. These values have been set after a thorough analysis of the network and some tests to correct errors. However, as variable choice depends on the data collection conditions, this sensitive step must be repeated if the database changes. The method therefore does not allow for easy replication of the research on another bus line or network. One of the objectives of this paper is to make signals filtering easily replicable. Hence, we tried to apply a clustering method so that this step does not depend on the survey context. The K-means algorithm, which allows to split observations of a dataset into k-predefined clusters was chosen for its simplicity and computation speed. The aim is to maximize homogeneity in the groups and heterogeneity between the groups (Cam and Neyman, 1967; Jain, 2010). For a set of points $x_1, x_2, \dots, x_n$, the algorithm is going to separate the n points in k clusters $S_1, S_2, \dots, S_k$, so that **equation 1** is fulfilled:

$$min \sum_{i=1}^{k} \sum_{x_j \epsilon S_i} d(x_j, \mu_i) \quad (1)$$

with μ_i the barycentre of the points in S_i.

As this method is only based on some variables computed for the signals, it is easily replicable to another bus line or another network. By using the scikit learn package of Python, the K-means algorithm is applied to a database whose unique identifier is the track (so each x_i is a track). Each track is assigned to a dedicated cluster. Then the same cluster is associated with every signal belonging to the track. The drawback of K-means algorithm is that we need to specify the number of clusters expected. Here, the number of ten clusters was determined mainly from a trial and-error-work. Ten groups might seem a lot, this choice was motivated by the fact that we prefer a lot of groups of well separated observations, instead of only two groups, one of passengers and one of parasites but with misplaced tracks. This implies to choose

afterward if a group is made of passenger or parasite signals, this step is explained below. Using K-means algorithm, we must choose one parameter (the number of clusters expected) against two, three or more when we set threshold values, making this method more replicable. Among all the variables presented in **table 1** we decided to use the six following for the clustering:

- Number of detections of the track
- Mean speed of the track
- Standard deviation of the distance to the barycenter of the track
- Standard deviation of the signal strength of the track
- Standard deviation of the time lapse between two consecutive signals of the track
- Standard deviation of the distance between two consecutive signals of the track

This choice was based on an observation of a small sample of tracks on which tracks that were highly likely to be emitted by passengers or, on the contrary tracks that were likely to be unwanted signals, were observed. Variables that were the most different for these two kinds of tracks were considered for the clustering. As an example, most of the observed tracks corresponding to bus passengers presented a high standard deviation of the distance from the barycenter of the track, whereas this value was low for most of the observed tracks corresponding to unwanted signals. Indeed, a low standard deviation of the distance from the barycenter of the track implies that all the points of the track are homogeneously distributed around the barycenter. On the contrary, when the track includes enough signals, the shape of a bus trip follows a line, more or less winding. This leads, most of the time, to distances from the barycenter rather heterogeneous, that is to say with a high standard deviation. This variable on its own cannot predict the fact that a track corresponds to a passenger of the bus or not because this observation is not obvious for short tracks. For the standard deviation of the signal strength, a high value implies that the signals of a track have been detected with heterogeneous strengths. The signal strength is very correlated with the distance between the sensor and the emitting object at the time of the detection. Hence if the strength is low, the object is probably far from the sensor, or there is something obstructing the emission. A track with signals detected with heterogeneous strengths could belong to an object whose position from the sensor is varying a lot, which is mainly not the case with a fixed sensor in a bus. Therefore, the six above mentioned variables seemed to be relevant for separating the unwanted signals and signals emitted by passengers. A study of the correlations between the variables completed the choice of the input variables for the K-means algorithm. As the variables used to build the clusters have very different ranges, they have been normalized before applying the K-means algorithm thanks to the "z-score" formula. Hence, for the variables mentioned above, each observation is transformed as in **equation 2**:

$$Z = \frac{(x-\mu)}{\sigma} \tag{2}$$

with μ the mean and σ the standard deviation of the sample. This leads to a sample with a mean at 0 and a variance of 1.

The objective is to identify the clusters which do not correspond to passengers of the bus. Each group formed by the K-means possesses singular properties, identifiable both by variables used for modeling and by illustrative variables. These are for example, the length, the duration, and the detection frequency of the track. According to their properties, some clusters are defined as passengers signal and others as unwanted signals. The latter were eliminated, and the Origin-Destination matrix is finally constructed from the passenger signals clusters.

4.2. Comparison to ground-truth data

In order to demonstrate the good quality of O-D matrices derived from Wi-Fi data, they will be compared to real data. Depending on the data source, several parameters will be observed to determine if Wi-Fi data represent mobility of bus passengers as well as ground truth data. The share of boardings at each stop over the total boarding of the line, in both directions of the bus route, is the first parameter observed,

for the three data sources available (optical counts, smartcard data and O-D onboard survey). The alightings are watched the same way for the optical counts and the O-D survey. Then the relative bus load (with respect to the maximum bus load) all along the bus route is considered (Wi-Fi data are compared with optical counts and O-D survey). Finally, the share of passengers on each O-D pair is studied for Wi-Fi data and the O-D survey. For all these parameters, the trends are compared and for most of them, R-squared values are computed to show the good fit between the data sources. When the samples are the same for the different data sources, absolute values of boardings at each stop are observed.

5. Results

In this section, we will first present the results of the clustering applied to the dataset presented above. Details will be given on the clusters obtained, and on the characteristics that differentiate them from each other. Then, we will explain why some clusters were suspected of not containing signals from bus passengers and were removed. Finally, Origin-Destination flows derived from the clustering will be compared to ground truth data, i.e., O-D onboard survey, smartcard data and optical counts. As a remainder, the dataset used in this part only contains data collected on line T3 on the 11th of October 2018. All the results presented below concentrate on this dataset.

5.2. Clustering algorithm

By applying the K-means algorithm to our dataset, we get 10 clusters of detected tracks. Almost 70% of the total variance is explained by the intergroups variance, so the clustering method used separates the tracks quite well. Observations within a group are close to each other, whereas observations differ from one group to another. **Table 3** summarizes intra and inter-groups variances.

Table 3: Intra and inter-groups variances

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Tracks	5577	800	2432	26194	5430	3843	934	114	3939	2213
Variance intragroups	7302.6	5433.1	9727.3	17088.3	9522.1	9207.8	9425.9	5570.8	9677.3	8832.7
Variance intergroups	7092.1	18037.2	22532.3	31332.0	20734.5	19740.2	18406.4	22624.7	8939.3	35344.3
Variance intergroups/total variance	0.691									

Based on the characteristics of the different groups identified by the clustering (**table 4**), we can classify the tracks as coming from “passengers” or “non-passengers”. Tracks belonging to clusters 1 and 4 have an average speed of 7 km/h or less, which is too low for a bus which partly circulates on dedicated lanes. Cluster 3 and cluster 6 present very few detections and either the distance or the time lapse between two detections of a track is very irregular. This makes these signals rather unreliable. Cluster 5, besides having an average speed a little bit too high for a bus (more than 25 km/h) presents a very low standard deviation of the distance from the barycenter of the track (either the signals are all grouped, either they form a circle around the barycenter, as mentioned in the methodology) and a low number of detections. Finally, in cluster 7, the standard deviation of the signal strength is very high, suggesting that the signal is sometimes very close and other times very far from the sensor, thus not inside the bus. The four remaining clusters (2, 8, 9 and 10) seem to be composed of signals emitted from connected objects belonging to bus passengers. The average speed is between 11 and 15 km/h. The number of detections per track is average or high. The standard deviation of the signal strength is average. The standard deviation of the distance from barycenter is quite high, which lets us suppose they are ungrouped signals. Finally, the standard deviation of the

distance or time lapse between two consecutive signals is not very high, so we can exclude the fact that they are very irregular, hence unreliable.

Table 4: Main characteristics of each cluster

Clusters	Observations	Number of detections per track	Track speed (m/s)	Distance from track barycenter std deviation (m)	Track signal strength std deviation (dBa)	Distance from next signal std deviation (m)	Time lapse from next signal std deviation (s)
1	10.8%	8	2.02	24	1.66	1209	32582
2	1.6%	78	3.97	383	4.65	307	14011
3	4.7%	5	4.15	62	2.08	2508	70467
4	50.9%	8	1.34	13	1.45	60	9987
5	10.5%	6	7.50	30	1.47	294	22757
6	7.5%	6	2.96	45	1.80	528	165416
7	1.8%	11	2.82	105	13.46	469	31069
8	0.2%	240	4.09	611	5.51	190	7787
9	7.7%	13	4.37	240	3.99	656	35650
10	4.3%	16	4.74	616	5.11	876	41380
Clusters “passengers” (2,8,9,10)	13.7%	25	4	380	4.44	678	34546
Clusters “non-passengers” (1,3,4,5,6,7)	86.3%	8	3	24	1.80	416	31591
Total	51476	10	3	73	2.16	452	31997

Among all the variables computed, some were not used for the clustering. They are the track duration and distance travelled, the detection frequency, the share of the track duration over the MAC address total detection duration… Some characteristics also emerge from these variables. Clusters containing tracks with a quite long duration and distance traveled are considered as emitted by bus passengers. These two parameters are represented for each cluster in **figure 2**. In the cluster 9, tracks are shorter but still not considered as unwanted signals, as the share of the track duration over the MAC address total detection duration is high. It means that the connected object they come from did not emit a lot of other tracks (connected objects emitting too much can be considered as unreliable). **Figure 3** shows that clusters 2, 8, 9 and 10 (identified as “passengers”) have the highest share of track duration/MAC address total detection duration. This ratio is quite high for cluster 3 because it has, on average, a very short MAC address total detection duration.

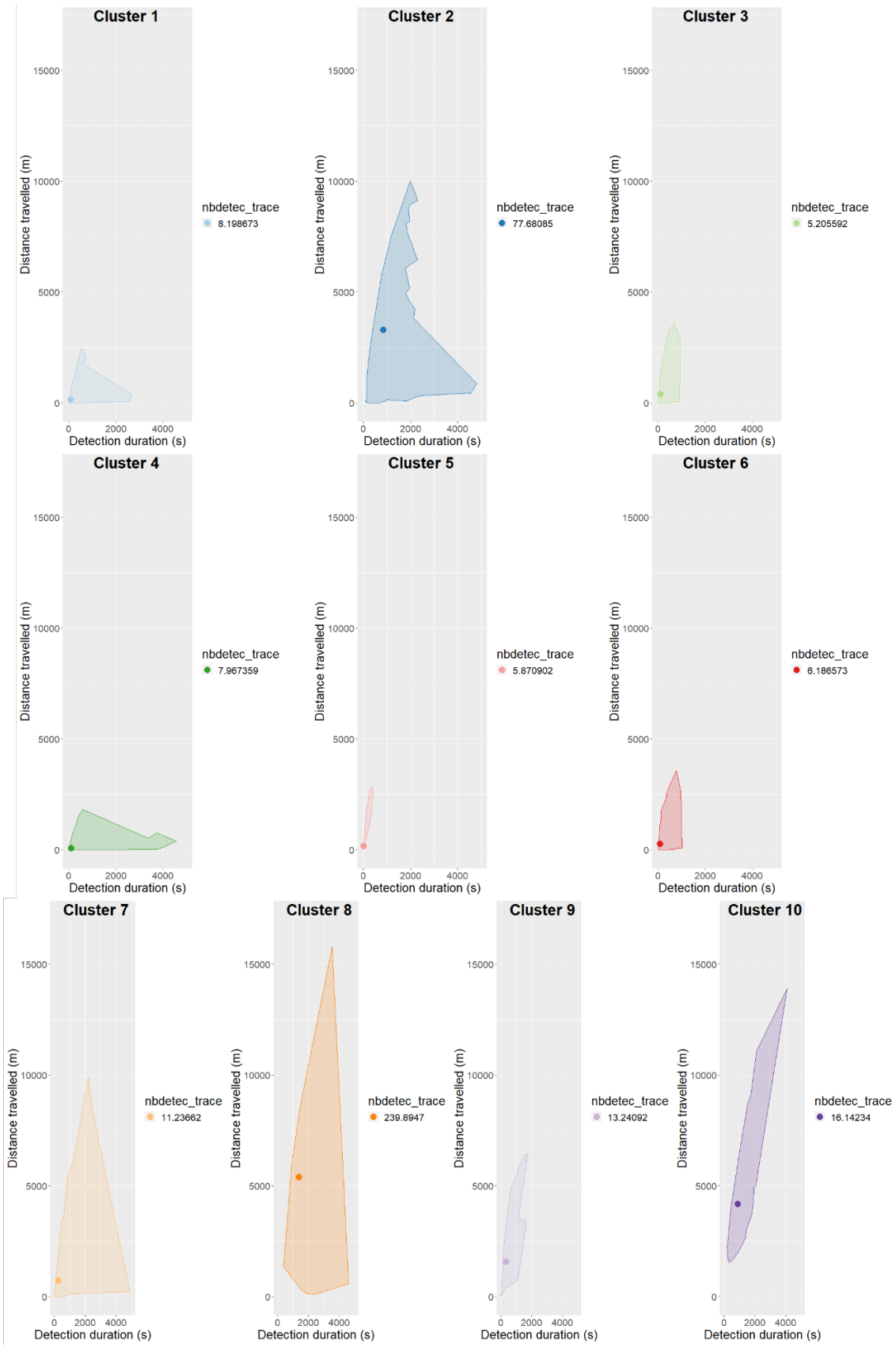


Figure 2: Concave hulls of mean duration and distances travelled by cluster

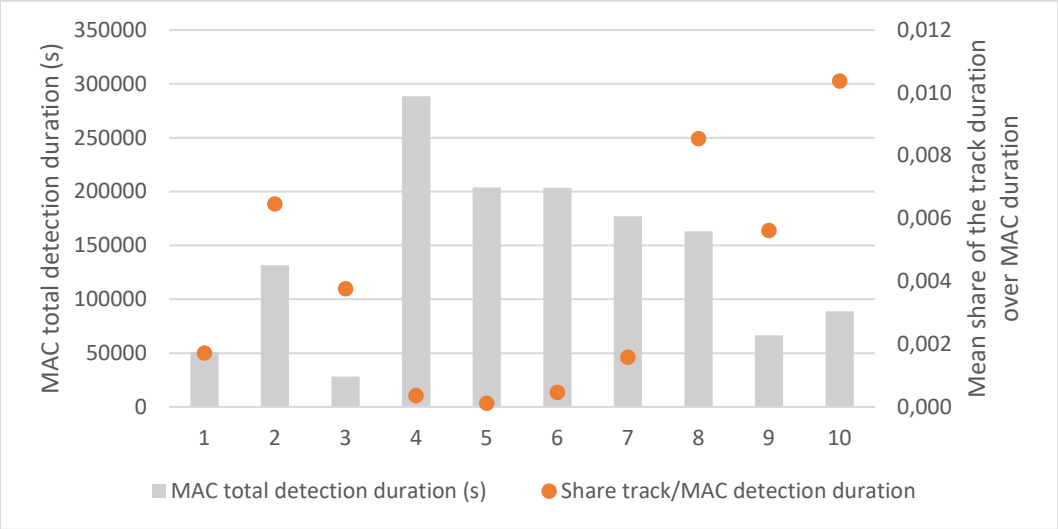


Figure 3: Mean track duration over MAC total detection duration by cluster

We can then build the Origin-Destination matrix from the tracks of the clusters 2, 8, 9 and 10 and compare it to ground-truth data, to determine if the methodology leads to qualitative results. An algorithm based on the association of the closest stop to the first and last detected signals from a track was used to build the O-D matrix. The selection of the four clusters identified above represents a total of 6 660 O-D pairs on line T3.

5.3. Comparison to ground-truth data

The quality of the data collected with Wi-Fi sensors is evaluated in two ways. First, we compare data from one box installed in a bus on line T3 on the 11th of October 2018, with two other data sources: optical counts collected by a camera and smartcard data from validations made in the same bus. Then, to evaluate the reliability of the method on a larger perimeter, we compare data from 16 boxes positioned in different buses travelling on line T3 on the same day, with data from the O-D survey and smartcard data. As the O-D survey and the smartcard data used in this case evaluate the traffic on the whole fleet, whereas "Laflowbox" sensors are positioned in 16 buses only, the absolute values of indicators computed with "Laflowbox" data cannot be compared to those of indicators computed from the smartcard data and the OD survey. However, indicators using relative values will be considered. In the following sections, only the Wi-Fi signals belonging to clusters 2, 8, 9 and 10, identified as passengers, are used to build the O-D matrix and to identify the number of boardings and alightings. We tried to include cluster 3 and cluster 7 in the analysis, due to their relative proximity with other passengers' clusters. This solution was not retained, as it deteriorated the correspondence with ground-truth data.

5.2.1. Data from one box

On the 11th of October 2018, a box was placed in a bus on Line T3. In the same bus, a camera for optical counts was also installed. Hence, we can easily compare the results of three data sources, collected on the same day: Wi-Fi data, smartcard data and optical counts.

When looking at absolute values, "Laflowbox" detects a total of 559 passengers. In the same bus, we identify 2290 passengers thanks to optical counts, and 1602 by analyzing smart card records. Hence Wi-Fi sensors detect about 25% of the passengers identified by optical counts and 35% of those who validate their ticket.

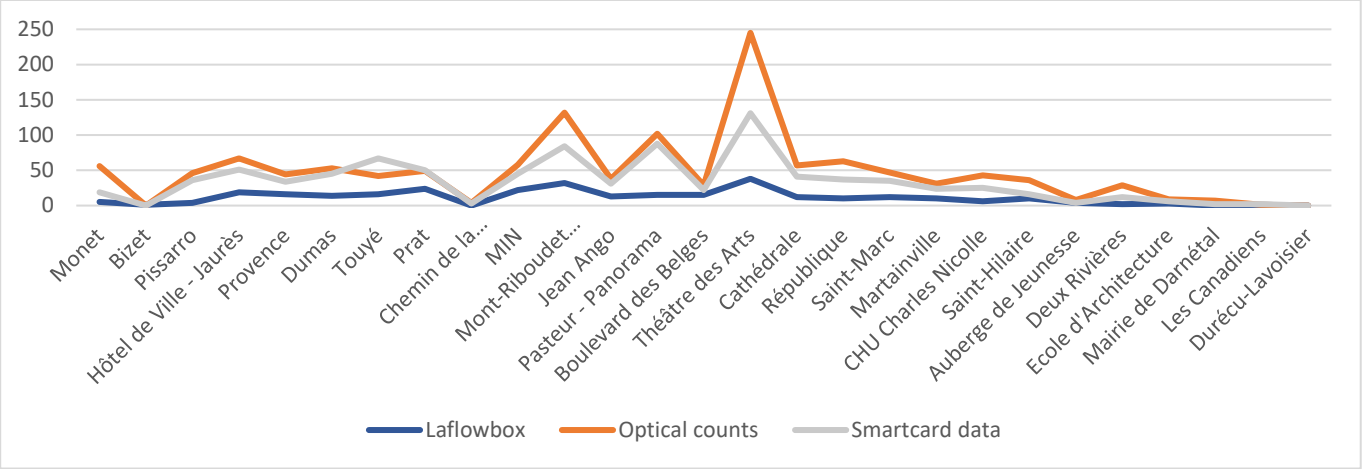


Figure 4: Absolute boardings at each stop – T3 direction 1 – 1 bus

For most of the stops, the absolute boardings follow the same trend, in smaller proportions for the Wi-Fi data and the smartcard data (**figure 4**). It is only at the stop "Théâtre des Arts" that the peak seems underestimated by the Wi-Fi data and the smartcard data. A good fit between the curves could be highlighted if we consider relative values. For direction 1, **figures 5 and 6** show the share of boardings and alightings (for Wi-Fi data and optical counts) at each stop over the total boardings/alightings of the bus. As the curves obtained for the boardings and alightings in direction 2 are similar, they are not presented here, the figures are in **annexes A and B**. The R-squared values between Wi-Fi data and optical counts are 0.70 for the boardings and 0.81 for the alightings in direction 1; 0.87 for the boardings and 0.73 for the alightings in direction 2. Between Wi-Fi data and smartcard data, the R-squared values are 0.79 in direction 1 and 0.81 in direction 2. All these values reflect a good fit between the three data sources. For comparative purposes, R-squared values between optical counts and smart-card data are 0.86 in direction 1 and 0.95 in direction 2.

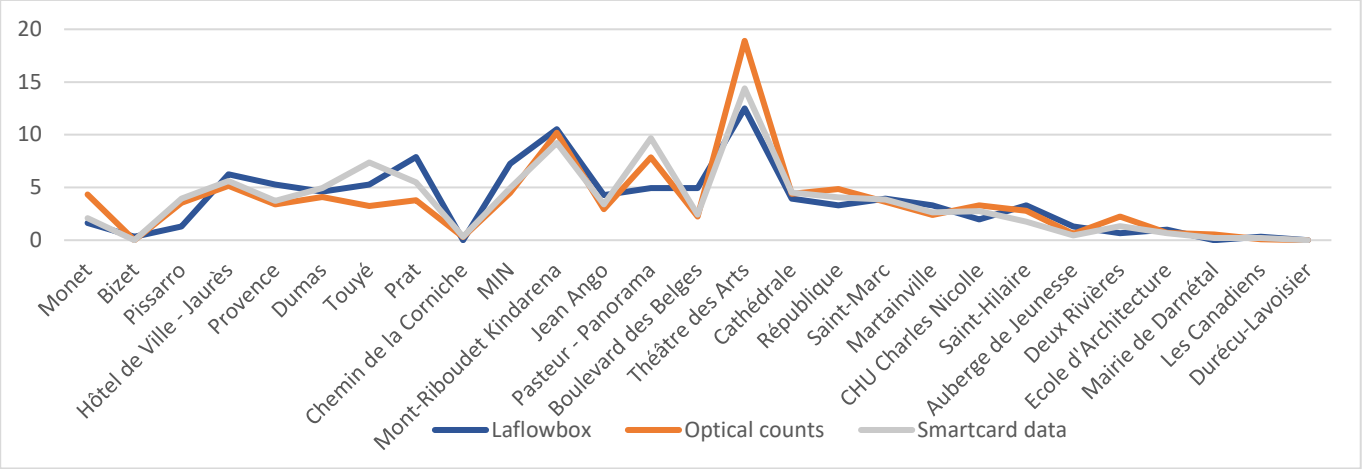


Figure 5: Relative boardings at each stop – T3 direction 1 – 1 bus

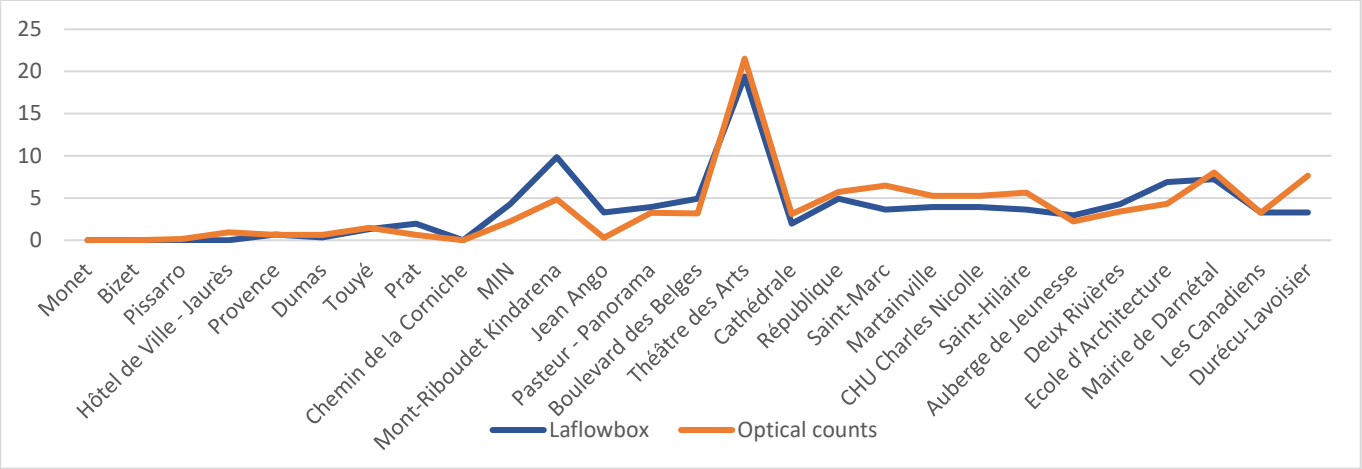


Figure 6: Relative alightings at each stop – T3 direction 1 – 1 bus

Finally relative bus loads (with respect to the maximum load observed) are represented in **figures 7 and 8**, respectively for direction 1 and 2. The fit of the curves is equal to 0.88 for direction 1 and 0.90 for direction 2. The bus load is only computable for Wi-Fi data and optical counts because smartcard data only gives us boarding information.

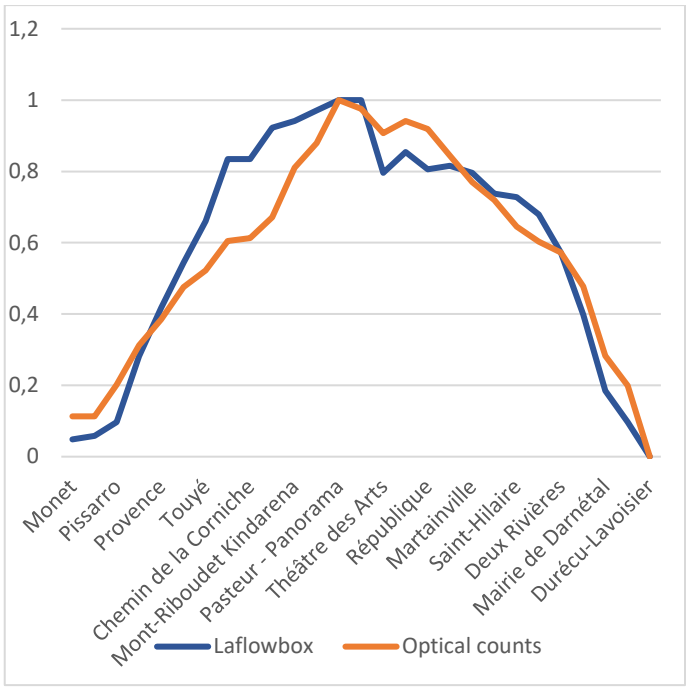


Figure 7: Relative bus load – T3 direction 1 – 1 bus

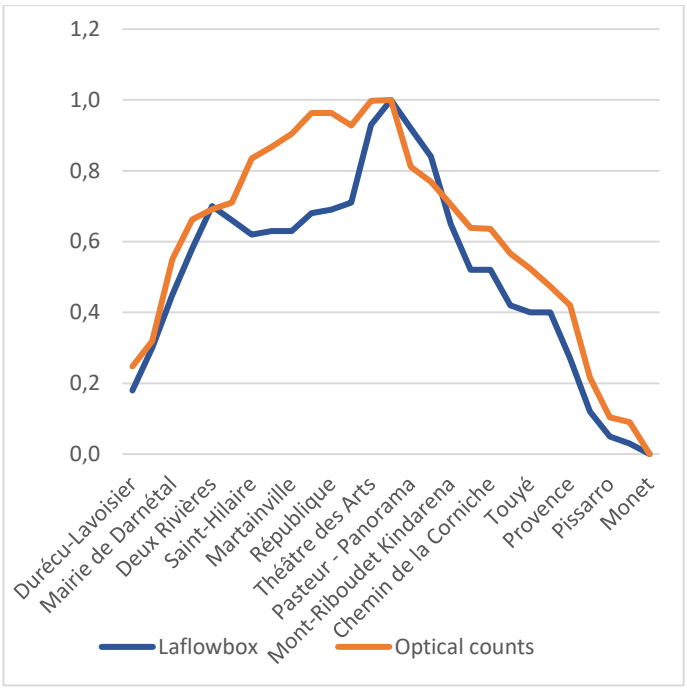


Figure 8: Relative bus load – T3 direction 2 – 1 bus

In this section, the Wi-Fi data collected in one bus, were compared with optical counts and smartcard data. Whatever the source, we only kept data coming from the bus in which were installed the Wi-Fi sensor "Laflowbox" and the camera used for optical counts, on the same day. The first comparison concerned the absolute number of boardings/alightings at each stop and proved that Wi-Fi data pictures the variation of the number of passengers along the line well. This conclusion is supported by the study of the share of boardings/alightings along the line. Indeed, the shares given by the three different data sources are

very similar, with a difference between smartcard data and Wi-Fi data not exceeding 4.7% and 6.4% between optical counts and Wi-Fi data.

5.2.2. Data from 16 boxes

In order to validate the reliability of the methodology on a larger scale, we can also analyze the data from all the 16 boxes running on the Line T3 on the 11th of October 2018. For comparison, we dispose, the same day, of an O-D onboard survey implemented on all vehicles and of smartcard data coming from all the buses circulating on line T3.

Comparing absolute values does not make sense here because Wi-Fi data are captured on 16 buses only whereas all the buses travelling on line T3 are represented with the O-D survey and the smartcard data. The total number of observations with Wi-Fi sensor is 6 660 for both directions, when 24 615 passengers are identified with the O-D survey and 11 492 thanks to smartcard data. The following analysis will focus on relative comparisons. Here again, we will look at the share of boardings and alightings at each stop over the total boardings/alightings, as well as the bus loads, from the Wi-Fi data, the O-D survey and the smartcard data. In order to validate the methodology on a larger scale, it is expected that these indicators are similar for the three different data sources.

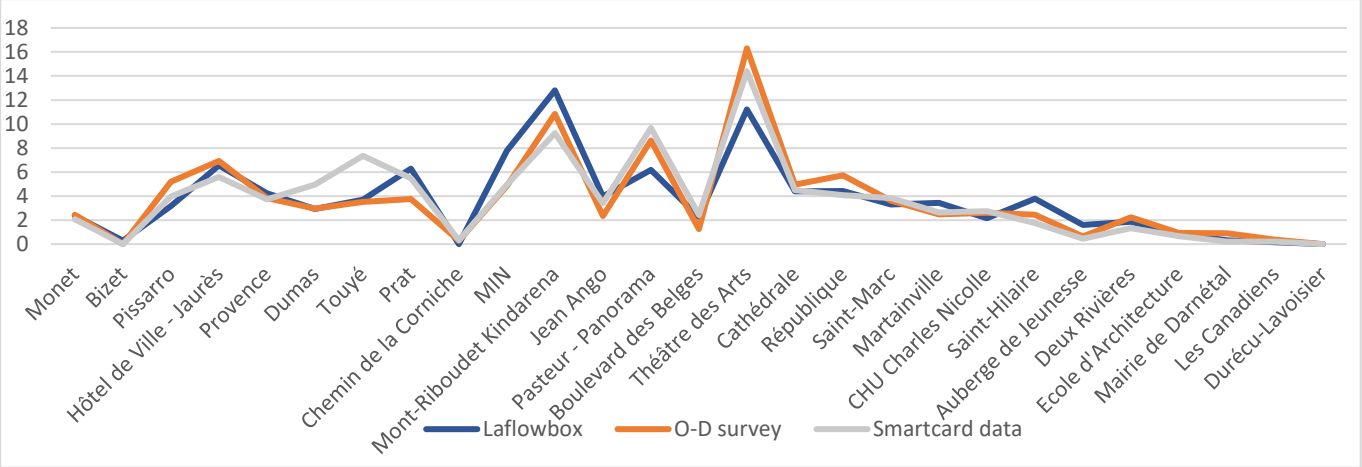


Figure 9: Relative boardings at each stop – T3 direction 1 – all vehicles

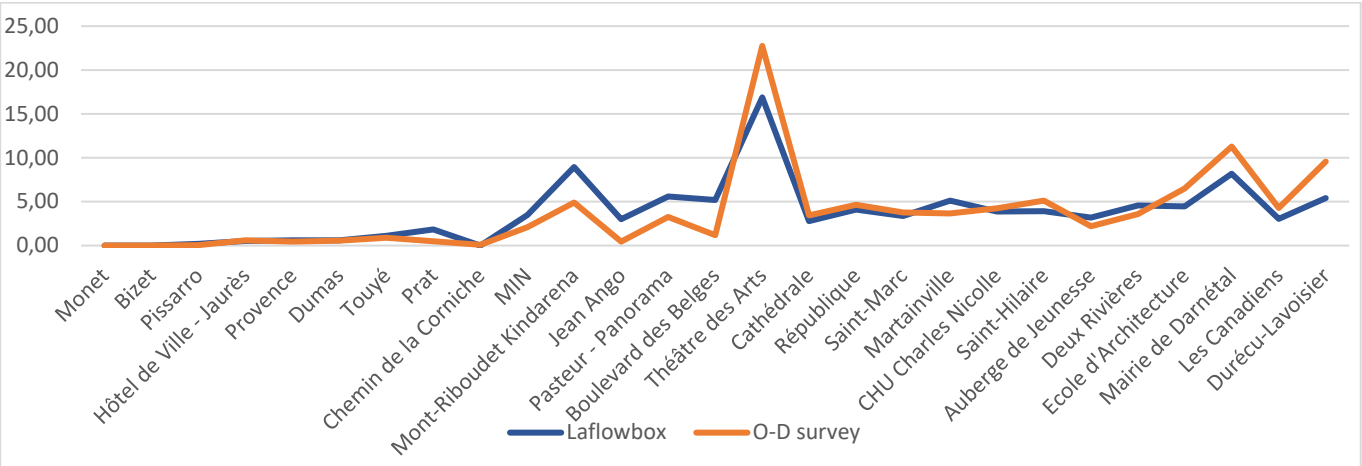


Figure 10: Relative alightings at each stop – T3 direction 1 – all vehicles

For direction 1, the share of boardings (respectively alightings) at each stop over the total boardings (respectively alightings) is quite similar between the different data sources (**figures 9 and 10)**. Here again, the same curves are obtained for direction 2 but not presented to alleviate the paper, they are in **annexes C and D**. The curves fit well, as shown by the R-squared values listed in **table 5**: all the R-squared values are above 0.80, whatever the direction or the data source (except for the alightings in direction 2 between the O-D survey and Wi-Fi data, mainly due to an underestimation of passengers at "Théâtre des Arts").

Table 5: R-squared values between the three data sources

Direction 1 / Direction 2	Laflowbox		O-D survey		Smartcard
	Boardings	Alightings	Boardings	Alightings	Boardings
Laflowbox	-	-	0.81	0.82	0.80
O-D survey	0.92	0.56	-	-	0.95
Smartcard	0.91	-	0.92	-	-

Concerning the bus loads (**figures 11 and 12**), the blue curve representing Wi-Fi data fits very well the orange curve (O-D survey): R-squared values are worth respectively 0.92 and 0.85 for direction 1 and 2. In direction 2 both curves are close even if the missing alightings at "Théâtre des Arts" and the overestimation at "Mont-Riboudet" lead to a bigger gap between the two curves on the second part of the line.

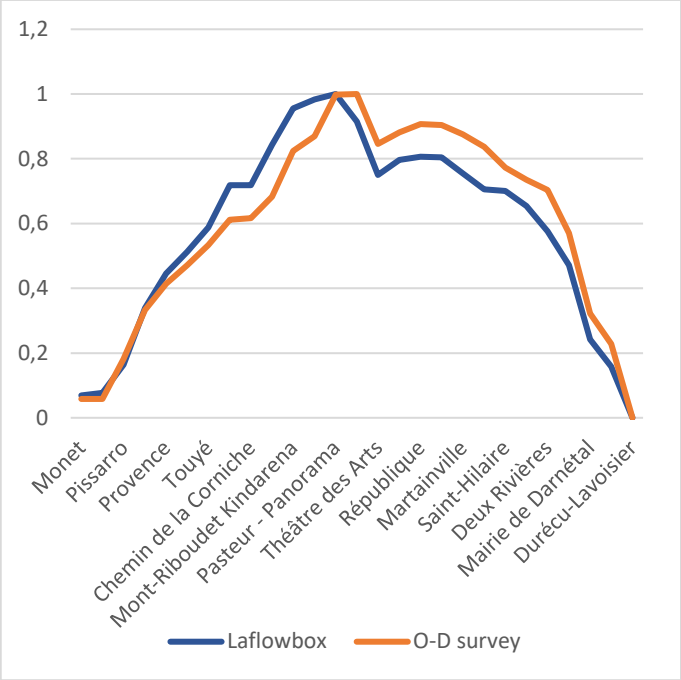


Figure 11: Relative bus load – T3 direction 1 – all vehicles

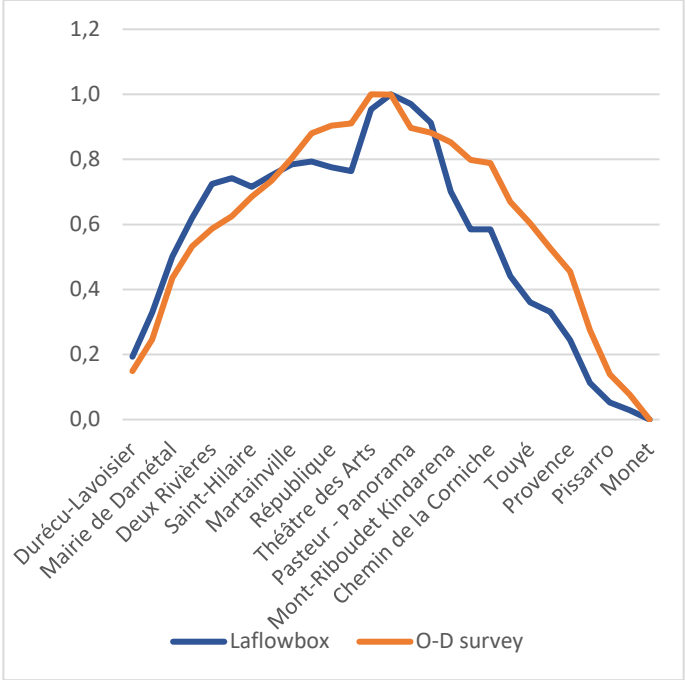


Figure 12: Relative bus load – T3 direction 2 – all vehicles

As we have seen in the previous analyzes, the share of boardings (or alightings) at stops is very similar from one data source to another, except at two or three stops. We wanted to see what kind of travels originating (or ending) at these stops are under or over estimated. So, to complete the previous analyzes, the difference between the shares of O-D pairs built from the O-D survey and Wi-Fi data is computed (**figure 13**). A blue box indicates that "Laflowbox" overestimates the pair whereas a red box indicates that

it underestimates the pair. It makes us suppose that the differences between both data sources mainly come from too many short O-D pairs and not enough long O-D pairs estimated with Wi-Fi data. For example, there is an underestimation of the long pairs originating or ending at “Théâtre des Arts” and an overestimation of the small pairs originating or ending in the zone “Mont-Riboudet” – “Boulevard des Belges”. For the other O-D pairs the differences are very weak.

OD survey - LFB (relative)	Monet	Bizet	Pissarro	Hôtel de Ville - Jaurès	Provence	Dumas	Touyé	Prat	Chemin de la Corniche	MIN	Mont-Riboudet Kindarena	Jean Ango	Pasteur - Panorama	Boulevard des Belges	Théâtre des Arts	Cathédrale	République	Saint-Marc	Martainville	CHU Charles Nicolle	Saint-Hilaire	Auberge de Jeunesse	Deux Rivières	Ecole d'Architecture	Mairie de Darnétal	Les Canadiens	Durécu-Lavoisier
Monet	0	0	0.2	0.2	-0	-0	-1	-2	0	-0	-0	-0	-0	-1	2.3	0.2	0.5	0.1	0.1	0.4	-0	0.1	0.1	0	-0	0	0.1
Bizet	0.6	0	0	-0	0	-0	0	0	0	0	0	0	0	0	-1	0	0	0	0	-0	0	0	0	0	0	0	0
Pissarro	0.4	1.1	0	1.3	0.2	0.3	0.9	-1	0	0.7	1	0.2	0.4	-0	5.2	0	0.5	0.4	0.4	-0	0.1	0	0.1	0	-0	0	0
Hôtel de Ville - Jaurès	1	2.2	2.7	0	-1	0.1	-1	-2	0.1	-1	-1	-0	0.2	-1	4.9	1.4	1.2	0.5	0.1	0.7	-0	-0	0.1	0	0.1	0	0.2
Provence	-0	0.5	1.9	0.8	0	-0	-1	-1	0	-2	-2	-1	-0	-1	3.6	0.1	0.6	0.4	-0	0.7	0	0.1	0.1	0	-0	-0	-0
Dumas	0.5	0.5	1.1	2.7	0.6	0	0.3	-0	0	-0	-0	-0	-0	-1	0.3	0.6	0.6	-0	0.4	0.9	-0	-0	0	0	0.1	0	0
Touyé	-0	0.2	0.2	1.4	0.4	0.5	0	-0	0.2	-2	-2	-1	0.4	-1	2.8	0.9	-0	0	0.4	0.4	0.3	-0	-0	-0	0.1	0.3	0.1
Prat	-1	-0	0.6	-1	-1	0.7	0.3	0	0	-2	-7	-2	-2	-1	1	0.1	0.3	-0	-0	0.4	0.1	0	0.3	0	0.2	0.1	0.1
Chemin de la Corniche	0	0.1	0.2	0	0	0	0.2	0.1	0	0	0.3	0	0	0	0.6	0.1	0	0	0	0.2	0	0	0	0	0	0	0
MIN	-0	0.2	1.2	0	-0	0.3	-0	-2	0.1	0	-9	-3	-2	-3	3.7	-0	0.5	0.4	-1	0.4	-0	-0	0.3	0	0.1	0	0.2
Mont-Riboudet Kindarena	1.4	0.8	1	3.1	-1	2.8	-1	-3	0.7	-6	0	-5	-6	-8	3.6	-0	0	0.6	-2	1.1	1.2	0.2	-0	1.3	2.3	1.1	2.2
Jean Ango	0	0.2	0.3	-0	-0	0.3	0.1	-1	0.2	-2	-3	0	-2	-3	-3	1.1	-0	-0	-0	-1	0.2	-0	0	0.2	0.1	-0	0.3
Pasteur - Panorama	-0	-0	0.5	0.8	0.7	1.8	-0	-2	0.2	-1	-11	-1	0	-1	6.1	0.4	1.6	1	-1	-1	0.8	0.2	0.4	0.7	2.4	1.3	0.6
Boulevard des Belges	0.1	0.2	0.5	-1	-0	1.8	-0	-1	0	-4	-7	-3	-0	0	0.4	-0	-2	-0	-1	-0	-0	-0	0.1	0.3	-1	-0	0
Théâtre des Arts	3.1	2.2	4.7	5.1	2.4	5	2.4	4.5	0.6	0.9	-11	-7	1.6	-0	0	-1	0.2	0.1	-2	0.5	4.5	-1	1.6	6.4	7.6	1.8	7.3
Cathédrale	1	0.8	0.7	2.3	0.4	0.5	0.4	0.2	0.1	0.3	-4	-2	-2	-0	-2	0	-0	-0	-1	-0	-0	-1	-1	2.5	2.9	-0	1.7
République	1.1	0.7	0.7	1.3	0.5	0.8	0.3	1.2	0.3	0.1	0.4	-1	0.9	-1	0.6	-1	0	0.2	0.1	-0	-1	-1	-1	3.2	2.7	0.6	3.2
Saint-Marc	0.4	-0	0.6	1.5	0.2	0.3	-0	0.3	0.1	0.9	0.4	-0	0.8	-0	2.6	-2	-1	0	0.2	-0	0.5	0.1	0.1	1	-0	0.4	0.3
Martainville	0.3	-0	0.4	0.5	0.1	0.4	0.4	0.4	0	0.2	0.4	-1	0.1	-1	1.2	-2	-0	-0	0	-1	0.4	-1	-2	-0	-1	0.2	0.2
CHU Charles Nicolle	0.5	-0	0.4	0.9	0.3	0.9	0.4	0.7	0.1	0.6	-0	0	-0	-1	1.7	0.3	-1	-1	-1	0	0.1	-0	-0	-1	2.4	0.3	0.8
Saint-Hilaire	0.1	0.1	0.2	0.3	0	0.1	0.2	0.6	0.1	-0	0.5	0.2	0.7	0.6	2.9	-1	-0	-2	-1	0.2	0	-1	-3	-2	-0	-0	-0
Auberge de Jeunesse	0.1	0	0.1	-0	-0	-0	0	0.1	0	0	0.1	-0	0.4	-0	0.5	-0	-1	-2	-1	-0	-2	0	-0	-2	-1	-0	-1
Deux Rivières	0.1	0	0	0.3	0.1	0.1	-0	-0	0	-0	-0	0.2	0.3	-1	1.6	-1	-0	-2	-1	-0	-3	-2	0	-1	-0	1.1	1.3
Ecole d'Architecture	0	0	0	0.1	0	0	0	-0	0	-0	0.6	-0	0.7	0.3	3.9	0.8	0.5	-2	-2	-3	-4	-2	-0	0	-1	0.3	0.2
Mairie de Darnétal	0.8	0	0.1	0	0.1	0	0.1	-0	0	0.3	0.2	-0	1.2	-1	8.7	0.6	2	-1	-1	0.8	-2	-2	0.2	-1	0	0.3	2.5
Les Canadiens	0.1	0.1	0	0	0	0	-0	0.1	0	0.1	0.3	-0	0.3	-0	3.3	-0	-0	-1	0.1	-0	-2	-2	-0	-3	-0	0	1.1
Durécu-Lavoisier	0.2	0.1	0	0.1	0	0	0	0.1	0	-0	0.9	0.1	0.6	0.2	4.6	-1	-0	0.2	-0	0.7	-2	-2	-2	-2	-2	0.4	0

Legend: : [-11‰ ; 9‰]

Figure 13: Difference of O-D pair shares between O-D survey and “Laflowbox” (‰)

In this section, we compared the data from all vehicles running on line T3 from: Wi-Fi signals, O-D onboard survey and smartcard data. Whatever the source, we gathered all the data we got from buses traveling on line T3 on the same day. Even if still not exhaustive, the number of Wi-Fi signals collected from “Laflowbox” is more important than in the previous section so we can test the reliability of the system. The relative number of boardings/alightings at each bus stop follow the same trend for the three data sources, as well as the relative bus loads (for Wi-Fi data and the onboard survey). Looking at the R-squared values confirms the similarity between the different sources, with a mean of 0.78 between Wi-Fi data and the onboard survey and 0.86 between Wi-Fi data and smartcard data. However, at some stops, a noticeable difference appears. The Wi-Fi sensor system seems to underestimate boardings and alightings at “Théâtre des Arts” and/or overestimate them at “Mont-Riboudet”. The observation of each O-D pair computed from

Wi-Fi data and the onboard survey suggests that these differences mainly come from an excess of small trips originating or ending at “Mont-Riboudet” (and the two or three neighbor stops). This probably causes the underestimation of the relative boardings and alighting at “Théâtre des Arts”, since its preponderance is “masked” by the exceeding number of passengers at “Mont-Riboudet”.

6. Discussion

The results presented in this paper show that Wi-Fi data reflects quite well the flow of passengers along the line. By using K-means algorithm, it is possible to automatically identify the Wi-Fi signals emitted by connected objects from bus passengers. The comparison of some key indicators (bus loads, share of boardings at each stop…) computed on a large sample of data from “Laflowbox” sensors with those from optical counts, smartcard data and onboard O-D survey confirms the quality of the data. No longer context-specific, this method could be easily replicable on other bus lines and other agglomerations.

However, some failures are observed. The comparison of each O-D pair computed from Wi-Fi data and the onboard survey suggests that these differences mainly come from an excess of small trips originating or ending in the central zone of the line. Several hypotheses can be mentioned to explain these differences. The first one concerns the sensors used to collect Wi-Fi data. “Laflowbox” is equipped with an antenna that offers great coverage. The sensor can detect a Wi-Fi signal coming from quite far, and wrongly consider it comes from a connected object belonging to a bus passenger. This is especially true when the bus follows a long straight line, where no building comes as a barrier to unwanted Wi-Fi signals. The central portion of the T3 line (between “Mont-Riboudet” and “Boulevard des Belges”) has this topography. This could be the origin of the higher number of passengers identified by Wi-Fi sensors at these stops. Besides, this long straight line is in the portion shared by the three lines T1, T2 and T3, maybe this can be a source of error for the Wi-Fi sensors too.

Another phenomenon concerning the onboard O-D survey can be added. These surveys tend to underestimate the number of small trips because people who make them stay a shorter time onboard and are therefore less likely to be interviewed. This could explain, at least in part, the difference of the number of small trips originating or ending in the zone “Mont-Riboudet” – “Boulevard des Belges”. Finally, we can mention that the onboard O-D survey is adjusted with optical counts of an historical mean day, not the reference day (11th of October 2018). This could also explain the gaps between the O-D survey and the other sources.

The reliability of smartcard data can also be discussed, as public transport tickets or subscriptions are subject to non-validation, either this action is conscious (fraud) or unconscious (forgetfulness). That’s why the absolute number of validations is lower than the optical counts. Most of the differences between Wi-Fi and smartcard data are observed between the stops “MIN” and “Cathédrale”, which are positioned on the busiest part of the bus line, where non-validation is most likely to occur.

To improve the quality of data gathered with Wi-Fi sensors like “Laflowbox”, some adjustments to the method could be suggested. For example, using an antenna with a shorter coverage, or placing several sensors in the same bus to limit the detection of parasitic signals, especially in the case of straight routes. However, this study shows that Wi-Fi data can already be considered as reliable and efficient to describe urban mobility behaviors. Using a clustering algorithm outperforms other filtering methods implying threshold values. By comparing mobility behavior estimated from Wi-Fi data collected in 16 buses to the one given by so-called ground truth data (optical counts, smartcard data and onboard O-D survey), we show that Wi-Fi data are qualitative and could be used in transport planning to estimate bus passenger mobility. Indeed, they describe bus passenger mobility in a similar way that smartcard data or onboard surveys do but have also significant advantages: Wi-Fi data can give information about origin and destination of the trip, which is not the case for smartcard data; they can be collected continuously, as opposed to the onboard O-D surveys that are conducted occasionally; and they require few material and human resources to be implemented.

7. Conclusions

Collecting mobility data is essential to understand the travel behavior of inhabitants and to ensure the sustainable development of transport infrastructures. In recent years, following the increasing cost of travel surveys and the development of technologies, new data sources are emerging. Among them, Wi-Fi sensors installed in city buses allow real and continuous collection of mobility data. This passive method requires no effort from passengers and represents a marginal cost for the public authority. Some experiences are presented in the literature, but the question of the quality of the data and the generalization of the system remains.

The main challenge when using Wi-Fi data to build Origin-Destination matrices lies in the selection of signals coming from actual passengers of the bus among the huge quantity of signals detected. This problem is, most of the time, addressed using threshold values. This does not allow for a consideration of spatial and temporal variabilities, so threshold values must be studied and set again when the setting changes. This challenge is answered by using the K-means algorithm, which allows to automatically identify the Wi-Fi signals of connected objects belonging to bus passengers and build an Origin-Destination matrix. This automatic process enables one to get rid of the step which consists of evaluating threshold values for every new setting and make the method replicable to other bus lines and other agglomerations. The second challenge is to verify the quality of the data collected and to guaranty the replicability of the method. By comparing "Laflowbox" data collected in one bus with other real data sources (optical counts collected by a camera and smartcard data from validations), this study showed the reliability of Wi-Fi data to understand mobility behaviors. Indeed, estimated O-D matrices are similar to the one given by ground-truth data, traditionally used in transport planning. Moreover, results remain stable if we compare Wi-Fi data collected in sixteen buses to real data commonly used in transport planning (O-D onboard survey and smartcard data from validations).

These two conclusions confirm the quality of mobility data gathered continuously in public transport by Wi-Fi sensors and validate the use of clustering algorithms to automatically get O-D matrices from these passive data. The most important result is probably the replicability of the method, which is free of temporal and spatial considerations. So, this work brings useful answers to transport operators and is valuable for public authorities, willing to collect massively mobility data at lower cost. Applying the method to the entire network (4 bus lines) or on another period could be relevant in future research, to confirm the replicability of the method. As we conclude that mobility behaviors inferred from Wi-Fi data is similar to that estimated by other data sources, future work could combine these datasets in order to make a powerful database for public policy makers.

CrediT authorship contribution statement and Declaration of Competing Interest

Léa Fabre: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Software. Caroline Bayart: Conceptualization, Methodology, Supervision, Writing – review & editing. Patrick Bonnel: Conceptualization, Methodology, Supervision, Writing – review & editing. Nicolas Mony: Conceptualization, Methodology, Software, Supervision.

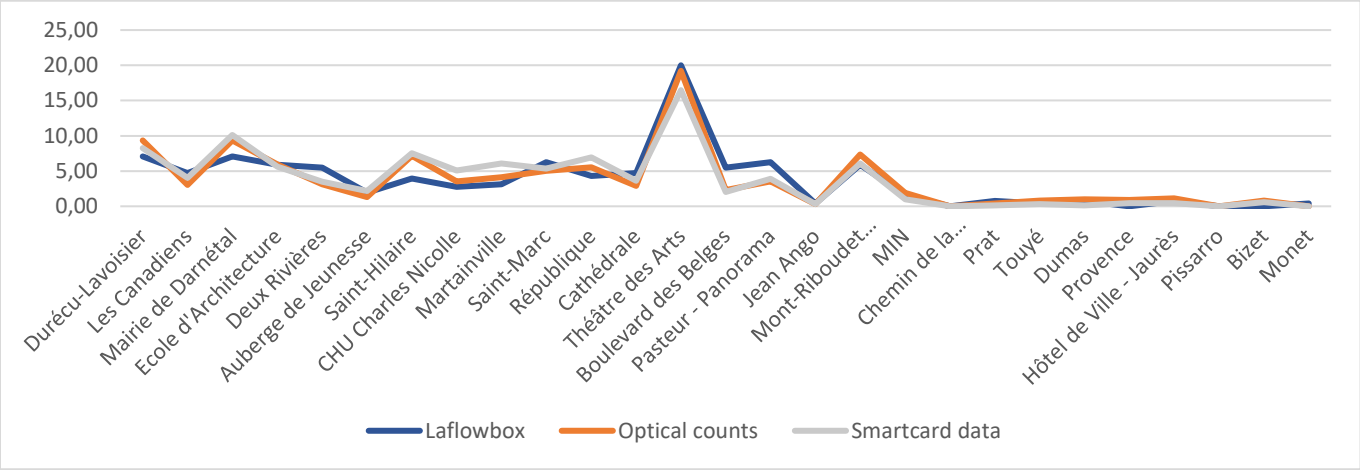
All the authors confirm that this work is original and has not been published elsewhere. It is not currently under consideration for publication elsewhere neither.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

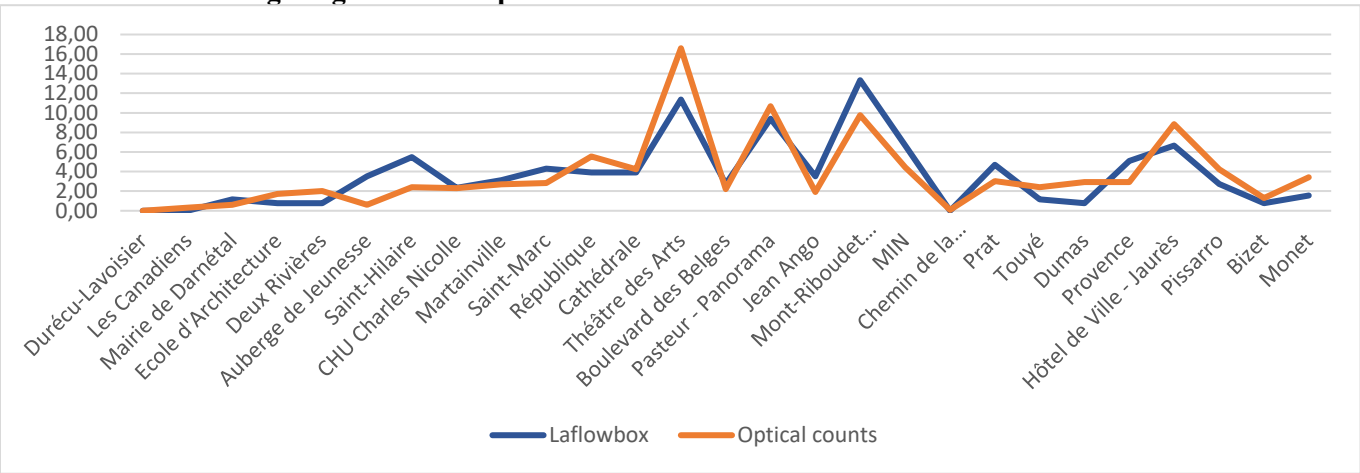
Acknowledgements

This research was conducted as part of a research agreement between Explain, the Urban Planning, Economics and Transport Laboratory (LAET) and the Actuarial and Financial Sciences Laboratory (LSAF). The authors acknowledge them for the financial support. The Métropole Rouen Normandie is also thanked for providing the data.

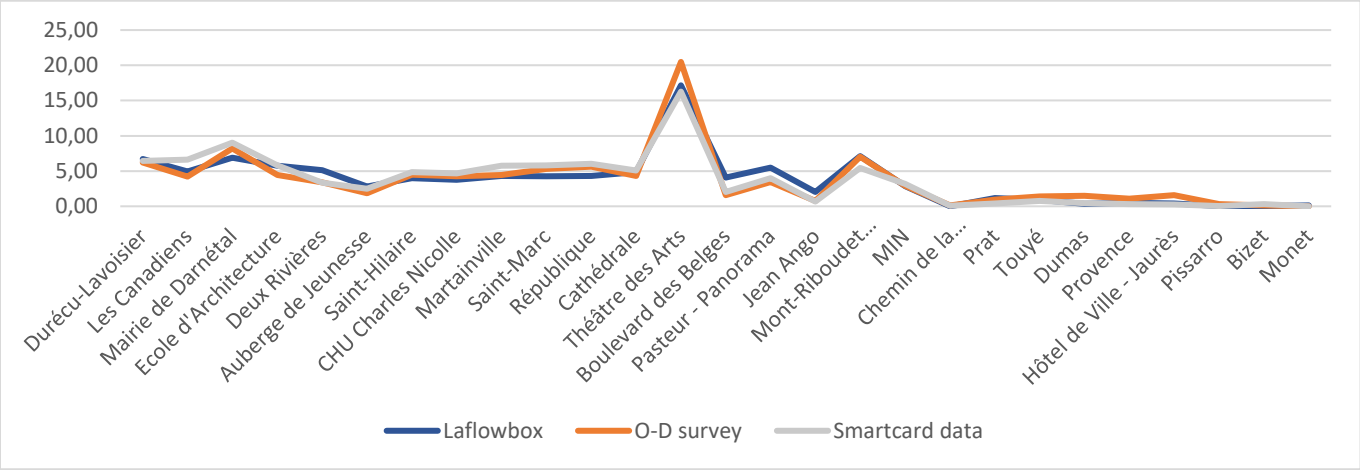
Annex A: Relative boardings at each stop – T3 direction 2 – 1 bus



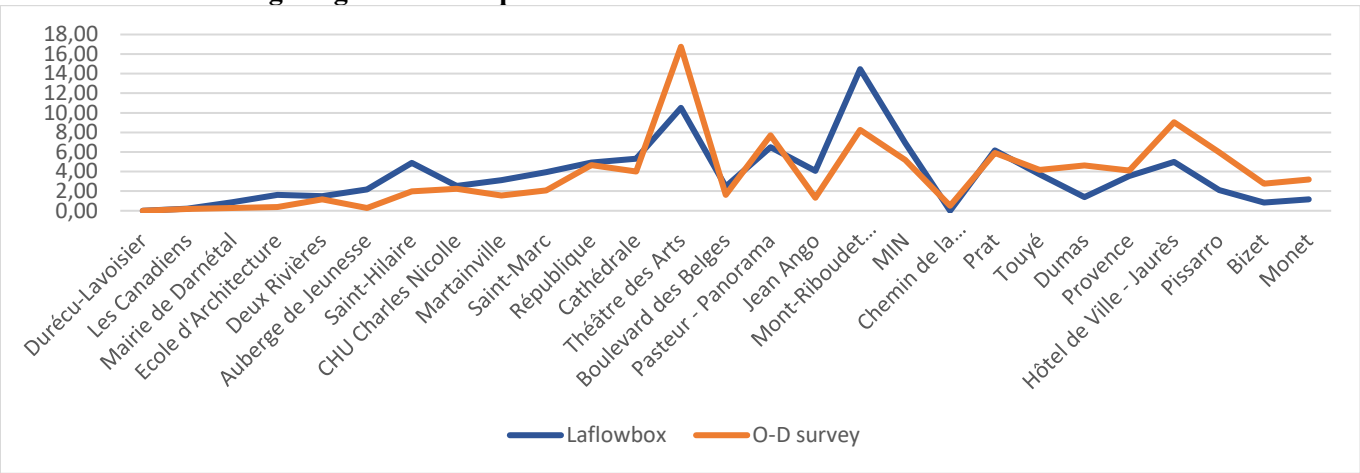
Annex B: Relative alightings at each stop – T3 direction 2 – 1 bus



Annex C: Relative boardings at each stop – T3 direction 2 – all vehicles



Annex D: Relative alightings at each stop – T3 direction 2 – all vehicles



References

Afshari, H.H., Jalali, S., Ghods, A.H., Raahemi, B., 2019. An Intelligent Traffic Management System Based on the Wi-Fi and Bluetooth Sensing and Data Clustering, in: Arai, K., Bhatia, R., Kapoor, S. (Eds.), Proceedings of the Future Technologies Conference (FTC) 2018, Advances in Intelligent Systems and Computing. Springer International Publishing, Cham, pp. 298–312. https://doi.org/10.1007/978-3-030-02686-8_24

Araghi, B.N., Christensen, L.T., Krishnan, R., Lahrmann, H., 2012. Application of Bluetooth Technology for Mode-Specific Travel Time Estimation on Arterial Roads, in: Proceedings from the Annual Transport Conference at Aalborg University. https://doi.org/10.5278/ojs.td.v1i1.5644

Axhausen, K.W., Zimmermann, A., Schönfelder, S., Rindsfüser, G., Haupt, T., 2002. Observing the rhythms of daily life: A six-week travel diary. Transportation 29, 95–124. https://doi.org/10.1023/A:1014247822322

Bayart, C., Bonnel, P., 2012. Combining web and face-to-face in travel surveys: comparability challenges? Transportation 39, 1147–1171. https://doi.org/10.1007/s11116-012-9393-x

Blogg, M., Semler, C., Hingorani, M., Troutbeck, R., 2010. Travel Time and Origin-Destination Data Collection using Bluetooth MAC Address Readers 15.

Bonnel, P., 2004. Prévoir la demande de transport.

Bonnel, P., Bayart, C., Smith, B., 2015. Workshop Synthesis: Comparing and Combining Survey Modes. Transp. Res. Procedia, Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia 11, 108–117. https://doi.org/10.1016/j.trpro.2015.12.010

Bonnel, P., Munizaga, M.A., 2018. Transport survey methods-in the era of big data facing new and old challenges. Transp. Res. Procedia 32, 1–15.

Borkowski, P., Jażdżewska-Gutta, M., Szmelter-Jarosz, A., 2021. Lockdowned: Everyday mobility changes in response to COVID-19. J. Transp. Geogr. 90, 102906. https://doi.org/10.1016/j.jtrangeo.2020.102906

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transp. Res. Part C Emerg. Technol. 26, 301–313. https://doi.org/10.1016/j.trc.2012.09.009

Cam, L.M.L., Neyman, J., 1967. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification. University of California Press.

CNIL, n.d. La loi Informatique et Libertés [WWW Document]. CNIL.fr. URL (accessed 10.25.21).

Deschaintres, E., 2018. Analyse de la variabilité individuelle d'utilisation du transport en commun à l'aide de données de cartes à puce. École Polytechnique de Montréal.

Dunlap, M., Li, Z., Henrickson, K., Wang, Y., 2016. Estimation of origin and destination information from Bluetooth and Wi-Fi sensing for transit. Transp. Res. Rec. 2595, 11–17. https://doi.org/10.3141/2595-02

Egu, O., Bonnel, P., 2020. Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon. Travel Behav. Soc. 19, 112–123. https://doi.org/10.1016/j.tbs.2019.12.003

Eisenmann, C., Nobis, C., Kolarova, V., Lenz, B., Winkler, C., 2021. Transport mode use during the COVID-19 lockdown period in Germany: The car became more important, public transport lost ground. Transp. Policy 103, 60–67. https://doi.org/10.1016/j.tranpol.2021.01.012

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.

Fukuda, D., Kobayashi, H., Nakanishi, W., Suga, Y., Sriroongvikrai, K., Choocharukul, K., 2017. Estimation of Paratransit Passenger Boarding/Alighting Locations Using Wi-Fi based Monitoring: Results of Field Testing in Krabi City, Thailand. J. East. Asia Soc. Transp. Stud. 12, 2151–2169. https://doi.org/10.11175/easts.12.2151

General Data Protection Regulation, n.d. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [WWW Document]. Eur-Lexeuropa. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504&qid=1532348683434 (accessed 10.25.21).

Hidayat, A., Terabe, S., Yaginuma, H., 2018. WiFi Scanner Technologies for Obtaining Travel Data about Circulator Bus Passengers: Case Study in Obuse, Nagano Prefecture, Japan. Transp. Res. Rec. 2672, 45–54. https://doi.org/10.1177/0361198118776153

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Ji, Y., Zhao, J., Zhang, Z., Du, Y., 2017. Estimating Bus Loads and OD Flows Using Location-Stamped Farebox and Wi-Fi Signal Data. J. Adv. Transp. 2017, 1–10. https://doi.org/10.1155/2017/6374858

Kurkcu, A., Ozbay, K., 2017. Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors: Transp. Res. Rec. https://doi.org/10.3141/2644-09

Kyritsis, D., 2017. The identification of road modality and occupancy patterns by Wi-Fi monitoring sensors as a way to support the “Smart Cities” concept: Application at the city centre of Dordrecht.

Malinovskiy, Y., Saunier, N., Wang, Y., 2012. Analysis of pedestrian travel with static bluetooth sensors. Transp. Res. Rec. 2299, 137–149. https://doi.org/Analysis of pedestrian travel with static bluetooth sensors

Mehmood, U., Moser, I., Jayaraman, P.P., Banerjee, A., 2019. Occupancy Estimation using WiFi: A Case Study for Counting Passengers on Busses, in: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT). Presented at the 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 165–170. https://doi.org/10.1109/WF-IoT.2019.8767350

Michau, G., Nantes, A., Chung, E., 2013. Towards the Retrieval of Accurate OD Matrices from Bluetooth Data: Lessons Learned from 2 Years of Data 12.

Mishalani, R.G., McCord, M.R., Reinhold, T., 2016. Use of Mobile Device Wireless Signals to Determine Transit Route-Level Passenger Origin–Destination Flows: Methodology and Empirical Evaluation. Transp. Res. Rec. 2544, 123–130. https://doi.org/10.3141/2544-14

Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N., Axhausen, K.W., 2015. Comparison of Travel Diaries Generated from Smartphone Data and Dedicated GPS Devices. Transp. Res. Procedia, Transport Survey Methods: Embracing Behavioural and Technological Changes Selected contributions from the 10th International Conference on Transport Survey Methods 16-21 November 2014, Leura, Australia 11, 227–241. https://doi.org/10.1016/j.trpro.2015.12.020

Myrvoll, T.A., Håkegård, J.E., Matsui, T., Septier, F., 2017. Counting public transport passenger using WiFi signatures of mobile devices, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). Presented at the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6. https://doi.org/10.1109/ITSC.2017.8317687

Namaki Araghi, B., Skoven Pedersen, K., Tørholm Christensen, L., Krishnan, R., Lahrmann, H., 2015. Accuracy of Travel Time Estimation Using Bluetooth Technology: Case Study Limfjord Tunnel Aalborg. Int. J. Intell. Transp. Syst. Res. 13, 166–191. https://doi.org/10.1007/s13177-014-0094-z

Nitti, M., Pinna, F., Pintor, L., Pilloni, V., Barabino, B., 2020. iABACUS: A Wi-Fi-Based Automatic Bus Passenger Counting System. Energies 13, 1446. https://doi.org/10.3390/en13061446

Oransirikul, T., Piumarta, I., Takada, H., 2019. Classifying Passenger and Non-passenger Signals in Public Transportation by Analysing Mobile Device Wi-Fi Activity. J. Inf. Process. 27, 25–32. https://doi.org/10.2197/ipsjjip.27.25

Patterson, Z., Fitzsimmons, K., 2016. DataMobile: Smartphone Travel Survey Experiment [WWW Document]. URL https://journals.sagepub.com/doi/abs/10.3141/2594-07 (accessed 7.29.21).

Pu, Z., Zhu, M., Li, W., Cui, Z., Guo, X., Wang, Y., 2021. Monitoring Public Transit Ridership Flow by Passively Sensing Wi-Fi and Bluetooth Mobile Devices. IEEE Internet Things J. 8, 474–486. https://doi.org/10.1109/JIOT.2020.3007373

Traunmueller, M., Johnson, N., Malik, A., Kontokosta, C.E., 2017. Digital Traces: Modeling Urban Mobility using Wifi Probe Data. Proc. 6th Int. Workshop Urban Comput. ACM KDD 2017 Halifax NS Can. 9.