



Origin–destination matrices from smartphone apps for bus networks

Benedetto Barabino¹ · Mauro Coni² · Massimo Di Francesco³  · Andrea Obino⁴ · Roberto Ventura¹

Accepted: 3 March 2024 / Published online: 10 July 2024
© The Author(s) 2024

Abstract

The knowledge of passenger flows between each origin–destination (OD) pair is a main requirement in public transport for service planning, design, operation, and monitoring, and is represented by OD matrices. Although they can be determined by traditional approaches (e.g., surveys, ride-check counts, and/or smartcard-based methods), the availability of new technologies and the proliferation of portable devices triggers an emerging interest in building OD matrices from the apps of bus operators. This research proposes the first framework for the estimation of OD matrices on transit networks by processing smartphone app call detail records (SACDRs). The framework is experimentally tested on a sample of 30 workdays of an Italian bus operator. The results are represented by easy-to-read control dashboards based on maps, which help quantify and visualise the OD matrices in the metropolitan area of Cagliari (Italy). The experimentation shows that the framework can properly estimate the number of trips for both origin and destination w.r.t. OD matrices built from household surveys: the mean absolute error is on average lower than five movements for 90% of the origins and 85% of the destinations.

Keywords Smartphone-app call detail records · APPs · Origin–destination matrix · Public transport

Abbreviations

| | |
|-----|---|
| AFC | Automatic fare collection |
| APC | Automatic passenger counting |
| AVL | Automatic vehicle location |
| BRT | Bus rapid transit |
| CDR | Call details record |
| DTO | Difference between consecutive timestamps |

This is an extended and revised version of a conference paper presented in *XXVI International Conference LWC 2023* in Brescia, 2023. An earlier version of this conference paper won the best younger researcher award for the best paper prize at the conference in the Public Transport: New Technologies and Economic Sustainability stream.

Extended author information available on the last page of the article

| | |
|-------|------------------------------------|
| GIS | Geographic information system |
| IPF | Iterative proportional fitting |
| ITS | Intelligent transportation system |
| MAC | Media access control |
| MCMC | Markov chain Monte Carlo |
| OD | Origin–destination |
| PTC | Public transport company |
| RFID | Radio-frequency identification |
| SACDR | Smartphone-app call details record |
| TAZ | Traffic analysis zone |

1 Introduction

Passenger flows are crucial components of the transit service for service planning, design (e.g. modifications and creation of routes), economical evaluation (e.g., cost benefit analysis), operation control (e.g. skip stopping and deadheading) and monitoring (e.g. checking a posteriori if a service is well provided), as well as the use of user-oriented measures of service effectiveness in public transport companies (PTCs) (Barabino et al. 2014; Olivo et al. 2019; Liu et al. 2022; Ventura et al. 2022; De Aloe et al. 2023).

The passenger flows between each origin–destination (OD) pair are described by OD matrices (e.g., Phithakkitnukoon et al. 2010). Their rows represent the possible origins of the trips and their columns the possible destinations. Each entry represents the number of passengers travelling from each origin to each destination in a considered timespan. In public transport, OD matrices can be built either at route or network level. At the route level, they describe passenger flows from boarding to alighting stops/stations on the route at hand. At the network level, stops or stations are clustered into traffic analysis zones (TAZs) and OD matrices describe flows between TAZs.

The estimation of OD passenger demand has been a challenging topic for several decades. It can be addressed by (i) on-board survey and/or ride-check count-based methods, (ii) smartcard-based methods and (iii) portable device-based methods.

On-board surveys are labor-intensive, time consuming, and expensive, whereas counting boarding and alighting passengers at each stop is easier and cheaper.¹ Moreover, on-board surveys could be people-biased and difficult to integrate with exogenous data sources, such as weather and traffic (Ge et al. 2021). Many methods were proposed to estimate OD passenger flows from these counts through a ‘seed’ matrix, which is usually developed from a classical survey (Simon and Furth 1985; Furth and Navick 1992; Navick and Furth 1994; Tamin 1997; Blum et al. 2010).

Recently, the spread of automatic vehicle location (AVL) and automatic fare collection (AFC) systems enabled the observation of the service and can be adopted for tracking vehicles and estimating passenger demand. These tools are endogenous

¹ The number of boarding and alighting passengers contains indirect information of OD, because at a bus stop, these counts totaled the sum of OD flows originating from and destined to that bus stop, respectively.

sources of large amounts of data and can support the generation of OD matrices using disaggregated data-driven methods (Zhang et al. 2007; Zhao et al. 2007; Trépanier et al. 2007; Rahbee 2008; Seaborn et al. 2009; Chu and Chapleau 2010; Wang et al. 2011; Munizaga and Palma 2012; Ge et al. 2021; Zúñiga et al. 2021). Although AFC systems return high-granularity data, none has been completely successful, because they do not account for fare evaders (Barabino et al. 2020, 2023). The recent proliferation of portable devices, such as mobile phones, smartphones, and tablets, presents new opportunities and challenges for collecting exogenous ridership data and tracking their movements throughout the network. Thus, data-driven methods such as Wi-Fi signals, large cell-phone data and apps have been proposed (Carrel et al. 2015; Chaudhary et al. 2016; Demissie et al. 2016; Mishalani et al. 2016; Håkegård et al. 2018; Tu et al. 2019; Nitti et al. 2020; Jee et al. 2023). Research on Wi-Fi signals has focused on detecting passengers (Håkegård et al. 2018; Tu et al. 2019) and counting on-board passengers (Mishalani et al. 2016; Nitti et al. 2020). Moreover, other studies have focused on determining passenger demand (by mobile phone call records of telco-operators, Demissie et al. 2016) or travel mode (by tracking voluntary passengers using customized apps, Carrel et al. 2015; Chaudhary et al. 2016). However, the necessity of gathering data from telco-operators and the need for passenger consent may limit the spread of these methods. For instance, Carrel et al. (2015) adopted an interesting app that helps voluntary passengers build a travel diary during their trips. However, the diary is too ‘sensible’ and, in the case in San Francisco city, only a low percentage of the demand was tracked according to APTA (2021).

Conversely, several PTCs nowadays implement their own apps that help passengers gather information about the estimated real arrival time of AVL-equipped vehicles at each stop/station (e.g. ATM in Milan and RATP in Paris). These apps can be queried by inputting the route and the stop/station of interest to gather information about the next real arrival time of a vehicle. Therefore, they are endogenous data sources and can offer insights into the trips of passengers at a low cost with limited privacy issues when passengers are not tracked. However, to the best of our knowledge, no study has explored the use of smartphone-app call detail records (SACDRs) data to build OD matrices in public transport.

This study aims to cover this gap. Specifically, this study presents a framework to infer OD movements and derive the corresponding matrix at the network level. The framework is experimentally tested using 9.6+ millions SACDR data provided by a bus operator, to demonstrate its viability in a real case study.

This study aims to contribute to both theory and practice. From a methodological perspective, the framework sheds light on an emerging research area, which has not been fully explored and presents several intermediate algorithms to handle SACDR data and build OD matrices. In addition, scholars could benefit from this study, as it offers a new way for estimating flows in public transport using emerging technologies. From a practical perspective, this study provides a different tool for passenger demand estimation in public transport.

The remaining paper is organised as follows. Section 2 summarises the relevant literature on OD matrices in public transportation. Section 3 presents the proposed framework for building an OD matrix. Section 4 presents the experimental results

for the overall network of a medium-sized Italian PTC. Finally, Sect. 5 presents the conclusions and research perspectives.

2 Literature review

2.1 Survey-based and/or ride-check methods

The classical way for building OD matrices usually consists of surveys on passengers and/or boarding and alighting count data by manual ride checks or APC.

OD matrices can be estimated at the network level for all transport modes by classical household travel surveys. Next, the trip distribution and modal choice models help estimate OD transit matrices (Ortuzar and Willumsen 2011). Other studies adjusted the estimation of OD transit matrices by boarding and alighting counts (Tamin 1997; Blum et al. 2010). For instance, Blum et al. (2010) estimated passenger demand by incorporating both household travel surveys and boarding–alighting data by a hybrid heuristic algorithm.

The OD matrices can also be estimated at the trip level. In this case, one can perform on-board travel surveys on a representative sample of transit rides and/or use ride check surveys (i.e., data on boarding and alighting passengers) on routes to build a new matrix or update existing matrices. Ride checks are easier to implement than on-board surveys. Transit trip surveys may help derive an a priori ('seed' or base) matrix. Next, data on boarding and alighting passengers are used to adjust, expand, and generate full matrices according to methods such as distance-based, bi-proportional, and similar iterative methods (Ben-Akiva et al. 1985; Furth and Navick 1992; Navick and Furth 1994; McCord et al. 2010; Mishalani et al. 2011). For instance, the iterative proportional fitting (IPF) method updates a seed OD matrix until the marginal row and column totals of the updated OD matrix satisfy the given boarding and alighting counts, respectively. Moreover, the IPF is straightforward to implement, computationally efficient and performs well in empirical studies.

Other methods use recursive approaches that do not require a seed (on-board) matrix, which can be inferred by boarding and alighting data only. However, a null base matrix is given. A null base implies that each feasible OD is equally likely to be travelled by the passenger (Simon and Furth 1985; McCord et al. 2010; Mishalani et al. 2011; Li and Cassidy 2007). For instance, Simon and Furth (1985) concluded that, despite this method being applicable for estimating route OD matrices for existing routes, special attention should be paid in the case of complex routes. Nevertheless, as in the case of a seed matrix, the IPF may be adopted to iteratively generate an OD-improved matrix.

Other research refined the methods for estimating the OD matrix using more complex algorithms and statistical properties. Specifically, recent studies considered the distribution of boarding and alighting data. Hazelton (2010) developed a Markov chain Monte Carlo sampler to infer OD movement matrices based on passenger counts. Ji et al. (2015a) proposed a method to recursively generate OD movement matrices from the first alighting stop to the last stop of a bus route using a Gibbs sampler-based Markov chain Monte Carlo method. Ji et al. (2015b) proposed an

expectation maximisation (EM) model that incorporated the OD matrix of the probability of trips from APC data and on-board OD flow survey data over the feasible OD flow matrices for each bus trip satisfying the APC counts. The heuristic solution to the model offered a better estimation than the IPF method when a null base or a poor on-board survey base were available. Moreover, the estimates are of similar quality to those of IPF when a large sample is adopted.

Most of these studies concluded that the integration of APC data with on-board survey information results in a better OD matrix estimation than that derived from an on-board survey alone or using a null matrix. However, even if the classical survey-based modelling is adopted, it presents some drawbacks: (1) Extensive classical surveys are undertaken every 5–10 years because they are costly and require laborious data processing and frequent updating. Hence, a priori OD matrices are probably outdated and not eagerly adopted by PTCs, because they cannot incorporate operational and specific considerations for some case studies. (2) Travel surveys usually suffer from large imprecisions in terms of coverage, spatial and temporal scale (Furth and Navick 1992). An example is the case of nonrespondent passengers following patterns that differ from those responding (e.g., standing passengers and short-trip passengers). (3) Methods using boarding and alighting passenger data require at least one seed matrix (a predefined or a null matrix) for the initialization.

2.2 Smartcard-based methods

A more recent way leverages automated data collection to build OD matrices without a seed or null matrix. The collection of high-resolution disaggregated data on passengers may be performed by AFC systems (Pelletier et al. 2011). They record the number of smart cards (i.e., tickets and/or passes) validated at specific points of the route either off board (e.g., at the gates of subways) or on board (e.g., in ticket machines on board buses). AFC systems are advantageous when radio frequency identification (RFID) technology is incorporated in the tickets (Rossetti and Turitto 2000; Oberli et al. 2010; Gonzalez et al. 2020) or in fully gated transit systems, because in these cases the origin and destination of each passenger are recorded.

However, AFC systems have some drawbacks. First, in most non-fully gated transit systems, data on transfer stops are not available because passengers are required to tap in only at the departure bus stop. Second, in many worldwide transit systems, pass holders are not required to tap in/out their tickets. Therefore, the incompleteness of stop/station information in such systems makes the trip determination a challenging issue. Third, AFC systems often coexist with other forms of tickets such as paper tickets and can result in incomplete data.

Other studies combined AFC data and AVL and/or GPS data and sometimes used APC for validation, to better infer boarding and alighting locations and times, thus providing a more accurate estimation of OD matrices (Zhao et al. 2007; Seaborn et al. 2009; Wang et al. 2011; Munizaga and Palma 2012). For instance, Zhao et al. (2007) developed a method for inferring passenger trip OD matrices in an integrated rail-bus transit system. This method combined AFC data to estimate the boarding-only of passengers and AVL data to determine the location of buses. Moreover,

owing to the only use of boarding information, they assumed that the stop where passengers board on a trip was where they alighted on the previous trip. Other studies on rail systems combined AFC with GIS data to estimate passenger demand (Rahbee 2008; Chu and Chapleau 2010).

Although all these methods are very valuable, some passengers may evade the fare because they may not carry the cards, buy tickets, or have invalid ones (Barabino et al. 2020). Although Munizaga et al. (2020) proposed a framework to correct errors in OD matrices due to fare evasion, its application worldwide would be tricky. Indeed, reliable smart card data require: (i) full smartcard transactions, where no other means of payment exist; (ii) a mode that has no evasion so that it does not have to be corrected, which is very uncommon; (iii) an OD survey conducted specifically on the zero-evasion mode, which can be used to correct partial fare evasion, even if this is less likely.

2.3 Portable device-based methods

The third way is emerging and rapidly evolving, as it is based on mobile devices (e.g., smartphones and tablets) as exogenous data sources. It outperforms both previous methods in terms of investment and maintenance costs because PTCs can avoid surveying passengers and/or installing many sensors (e.g., counting sensors), as these costs are “switched” to passengers.

Portable device-based methods help indirectly collect the data of passengers, because these data are on their devices. Although some passengers may carry more than one device (or anything), this is not a strong limitation, because some adjustment factors can be calibrated to improve the accuracy of scaling for inferring disaggregated origins and destinations. According to several statistics on some international reports, modern mobile devices have become essential to people’s daily lives (Drosouli et al. 2021). They reported that 81% of the world’s population owns a smartphone, and smartphone adoption among adults aged older than 50 years has increased from 62% (2017) to 79% (2019); the number of smartphones in use is growing by 5.6% each year.

Three main ways were considered for the estimation of OD flows from portable devices: (1) large-scale cell phone networks, (2) Wi-Fi technology, and (3) apps.

The use of a large-scale cell phone helps collect data on both the origin and destination when the device is connected to the cellular network. This is achieved by exploring Call Detail Records (CDRs), that may include a call, which is made or received (both at the beginning and end of it); a short message, which is sent or received; or when the user is connected to the Internet (e.g. to browse the web), which is adopted in the estimation of individual mobility (Caceres et al. 2008; Gonzalez et al. 2008; Calabrese et al. 2011; Deville et al. 2014; Iqbal et al. 2014). Conversely, to the best of our knowledge, only Demissie et al. (2016) presented a methodology to infer the origin and destination in public transport. Their method extracted the relevant origins and destinations of inhabitants to build OD matrices by CDR data provided by telco-operators. Although the large-scale cell phone method uses anonymous locations to avoid privacy issues, it presents some drawbacks. First,

it requires the participation of the telco-operator owning the infrastructure and, thus, the data, to be successfully applied. Second, some passengers may travel without a mobile phone; therefore, this demand cannot be estimated.

The use of Wi-Fi is a recent method for inferring passenger demand because mobile devices may be tracked by sniffing Wi-Fi traffic. This method relies on the identification of the media access control (MAC) address associated with a Wi-Fi signal. Origin and destination may be inferred once the device is connected to the Wi-Fi network, even if the passenger does not use the network (Tu et al. 2019; Liu et al. 2013). Mishalani et al. (2016) noticed that aggregated OD movements of multiple buses had better estimations than the disaggregated ones. Håkegård et al. (2018) observed that the integration of statistical models to estimate trip-level OD movements can estimate passenger loads close to those computed by APC data. Tu et al. (2019) presented a system that can infer both the origin and destination of passengers by fusing the network events generated by Wi-Fi devices (activated by passengers), AFC system and bus GPS information. Finally, Nitti et al. (2020) proposed iABACUS—a Wi-Fi based system that tracks anonymously passengers throughout their journey on buses and can return a simple OD matrix—while addressing the randomization of MAC addresses recently introduced by Google, Apple, and Microsoft (Myrvoll et al. 2017).

Although leveraging Wi-Fi technology is an interesting approach for building OD matrices, it presents some disadvantages. First, passengers might have turned off the Wi-Fi to minimise battery consumption (Tu et al. 2019; Nitti et al. 2020). Therefore, passenger demand may be underestimated. Second, the Wi-Fi access points are increasing rapidly, and several buses are expected to be equipped in the future. However, the number of people connected to Wi-Fi networks remains low. For instance, in Dordrecht (the Netherlands), the number of people connected to Wi-Fi networks ranged from 31 to 49% (Kyritsis 2017).

Apps may represent emerging ways for estimating passenger demand because the movement of the passenger may be followed. Recent research has shown how app-based systems help collect passenger data: (i) by tracking the individual location of passengers with high frequency (Carrel et al. 2015), (ii) providing information on on-board passengers (Chaudhary et al. 2016) and (iii) combining types of transportation (Lu et al. 2017). Carrel et al. (2015) proposed a system capable of tracking the use of transit by passengers at a disaggregated level by matching location data from smartphone apps and AVL data. Moreover, they were able to identify the passengers off or on board. Conversely, participatory sensing was used by Chaudhary et al. (2016) and Lu et al. (2017). The former proposed a cost-effective method to collect data on crowding by using a specific app integrated with GPS. The latter tracked the overall passengers' journey using different transportation choices once the app was installed.

However, if smartphone apps are adopted for tracking passengers, they must consent to install the app and, if no incentive is provided, little data could be collected. Conversely, if apps provide services (e.g., consulting information such as the real arrival time of a transit vehicle at a selected stop/station), passengers may not be reluctant to install and use the app, thus resulting in a large amount of data collected. No study investigated OD movement estimation by apps.

2.4 Summary of the past literature

Tables 1, 2 and 3 provide a comprehensive overview of all studies on the three ways for building OD matrices. Each reference in the tables reports the following attributes: authors (year); type of data; location of study (city/country); methodology and a quick summary of content and conclusions. Each table is sorted by year.

2.5 Gaps in the literature

All studies provided evidence that OD movements in public transport can be derived in several ways, even though it may be at different levels of detail. However, the related literature indicates some gaps, which are summarised hereafter.

First, classical survey-based methods are expensive in terms of time and money and can result in large inaccuracies in both space and time scales.

Second, smartcard-based methods result in more accurate data. However, support, maintenance, recharge network infrastructure and sensors in the vehicle may involve costs that are not negligible for PTCs. For instance, RFID technology incorporated in smart cards could be an interesting and viable solution, as shown by Oberli et al. (2010) and Gonzalez et al. (2020). However, hardware (e.g., antennas, controller system, sensors) should be installed on board in each vehicle and may be too expensive for PTCs. In addition, fare evasion cannot be detected.

Third, mobile device methods using large-scale cell phones require the participation of telephonic operators to collect data; thus, they can generate a high coverage error. Wi-Fi networks are not yet widespread in public transport.

Therefore, it is of interest to propose new methods for inferring OD matrices by leveraging the capability of apps. This study proposes the first method in this research area.

3 Methodological framework

In this section, a framework is presented to prepare and screen SACDRs, reconstruct the journey of passengers, infer the destination, build the OD matrix, represent detailed results over all TAZs and periods, and validate the estimates. Specifically, SACDRs data are processed to return an OD matrix at the network (TAZ) level. The rows of the matrix represent the origin TAZs, the columns the destination TAZs and each entry contains the number of trips for each origin–destination pair. The framework is summarised in the flowchart shown in Fig. 1, that adopts the notation recommended by the American National Standards Institute (ANSI) (Chapin 1970). The procedure is split in several steps, denoted by a dashed line, and are described in what follows.

Table 1 Studies with on-board survey and/or count-based ride-checks for OD matrices

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-------------------|---|---------------------------------|--|--|
| Liu et al. (2022) | AVL & APC | Hangzhou City (China) | Spatial analysis model | A real-time system for vehicle positioning and passenger flow sensing, to enable the collection and analysis of high-frequency data on both the operational status of BRTs and passenger flows at stops. An analysis is performed on the delay at intersections, the operational status, passenger flows and stranded passengers in the BRT system |
| Ji et al. (2015a) | On-board OD survey & APC | Ohio State University (USA) | Algorithms for expectation maximization method by APC and on-board OD flow survey data | A method to estimate transit route passenger OD flow matrices for time periods. When enough automatic boarding and alighting data are available, the method returns equal and/or better results than the IPF both in the case of large or small-size on-board surveys |
| Ji et al. (2015b) | Large-scale of On-board OD survey & APC | Ohio State University (USA) | MCMC | A sampler recursively generating OD flow matrices (without enumeration) from the initial alighting stop to the final stop of the considered bus route. The effectiveness of the MCMC method is shown as opposed to an existing MCMC one and the current practices |

Table 1 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-------------------------|--|--------------------------------------|--|---|
| Mishalani et al. (2011) | On-board OD survey & APC | Ohio State University (USA) | IPF | Examining the advantages of estimating route-level passenger OD flows by combining APC data and on-board OD survey data. The approach provides more precise OD flow estimates w.r.t. APC data alone: a larger on-board survey sample enhances the accuracy of these estimates |
| Blum et al. (2010) | On-board OD survey and boarding-alighting data | Greater Mumbai (India) | Algorithms estimating direct trips, transfer trips and altering the set of no-transfer trips | HDE algorithm estimating public transport demand at the network level. It produces a demand estimation with less than 1% deviation from both the boarding-alighting data and household survey |
| Hazelton (2010) | Boarding and alighting passengers | San Francisco (USA) | Bayesian approach & MCMC | A sampler for inferring OD traffic rates within transit systems. This sampler operates without requiring an explicit specification of the feasible trip set |
| McCord et al. (2010) | APC | Los Angeles County California, (USA) | IPF | Evaluation of the performance of the IPF for route-level OD flow determination with boarding and alighting data only. The IPF null procedure shows a good performance |

Table 1 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-------------------------|--|---------------------------------|---|--|
| Li and Cassidy (2007) | Boarding and alighting passengers | San Francisco (USA) | Algorithms for OD matrix for the vehicle trip and passenger alighting probabilities | An algorithm for transit passenger OD estimation generating estimates from the counts of passenger boarding and alighting at each stop along the route. The algorithm is computationally efficient w.r.t. the balancing method of OD estimation |
| Tamin (1997) | On-board OD survey and boarding–alighting data | Bandung (Indonesia) | GM combined with MNL and NLLS estimation method | Methods and techniques for modeling the public transport demand at network level. The required number of observed traffic counts are at least as many as the number of parameters: the more the link flows, the faster the convergence of the estimation method, the more accurate the estimated OD matrix |
| Navick and Furth (1994) | Boarding and alighting passengers | Boston and Miami (USA) | IPF and GM | A method for creating a complete OD matrix from the “seed matrix” by a function modelling travel propensity as a function of the travel distance |
| Furth and Navick (1992) | Boarding and alighting passengers | Southern California (USA) | IPF and Tsygalnitsky’s algorithm | A comparison between the recursive method of Tsygalnitsky and the IPF. The recursive method is not suitable in case of strong route competition and advises a correction procedure for aggregated on–off data at the segment level |

Table 1 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-------------------------|-----------------------------------|---------------------------------|--|---|
| Ben-Akiva et al. (1985) | On-board OD survey | Boston metropolitan area (USA) | IPF, constrained generalized least-squares, constrained maximum likelihood | A comparison between the current practice of expansion of OD matrices from on-board surveys and other methods. The IPF method is expected to enable transit operators to obtain much more accurate OD matrices and more reliable results at a small additional computational cost |
| Simon and Furth (1985) | Boarding and alighting passengers | Southern California (USA) | Tsygalnitsky's algorithm | Application of Tsygalnitsky's method to estimate OD flow. The method excels on simple routes and performs well on routes with some branching. However, passenger distribution becomes less fluid-like as the complexity of branching and the number of routes or competing transit modes increase |

A representative, but not a comprehensive list of references

APC Automatic Passenger Counting, AVL Automatic Vehicle Location, BRT Bus Rapid Transit, GM Gravity Model, HDE Hybrid Demand Estimation, IPF Iterative Proportional Fitting, MCMC Markov Chain Monte Carlo, MLR Multiple regression model, MNL Multi-Nominal-Logit, NLLS Non-Linear Least-Squares, OD Origin-Destination

Table 2 Studies on smartcard-based OD matrices

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|---------------------------|----------------------------------|---------------------------------|--|---|
| Zúñiga et al. (2021) | Historical travel time data, AFC | Valparaiso (Chile) | Probability distribution of travel destination choice, Bayesian algorithm, ANN | A methodology for real-time estimating, updating, and predicting OD matrices in short time intervals for a public transport corridor with validation at both the origin and destination points. The results are viable w.r.t. a static approach in real-world scenarios |
| Gonzalez et al. (2020) | RFID signals | Madrid (Spain) | Algorithms for calculation of individual journeys, the corresponding OD matrices, and revealing behaviors at each stop | A system identifying passengers while travelling on buses, while avoiding check-in and check-out actions. The results have high performance metrics: the individual journey estimation rates are larger than 87% |
| Munizaga and Palma (2012) | AFC & GPS | Santiago (Chile) | Algorithms for reconstructing the trip chain including alighting point estimation and modes | A methodology estimating public transport OD matrices using smartcard and GPS data for extensive multimode public transport systems. This approach estimates both the time and location of alighting passengers for over 80% of the boarding transactions |
| Wang et al. (2011) | AFC & AVL | London (UK) | Algorithms for inferring origin (matching AFC & AVL) and destination (implementing the trip-chaining method) | Potential applications of the automated data collection systems archived data to bus network planning with a focus on daily ridership variations and interchange time analysis. Alighting times for bus passengers are also estimated |

Table 2 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-------------------------|--------------|---------------------------------|--|---|
| Oberli et al. (2010) | RFID signals | Chile | Algorithms for inferring to recognize individual passengers (e.g., when they board and alight buses) | A lab experiment evaluating the performance of employing RFID for tracking passengers and determining OD matrices. The numbers of antennas, their position and the radiation pattern affect the recognition percentages of passengers |
| Seaborn et al. (2009) | AFC | London (UK) | Algorithms for transfer identification and linking journey stages by SQL | A methodology for inferring information to bus network planning. It shows that passenger flows among intersecting routes can be quantified by smart card data |
| Rahbee (2008) | AFC | Chicago (USA) | Algorithms for inferring destination, statistical sampling, path assignment, and network loading | A system estimating rail passenger flows for entry-only AFC data. Results mainly include estimations of travel direction, passenger volume and load profiles |
| Chu and Chapleau (2010) | AFC & GIS | Canada | Spatial statistics, spatial analysis with GIS, visualizations, and heuristic for data mining | A method to enhance the description of transit trips by incorporating multi-day elements. The outcomes show anchor points for each cardholder based on boarding patterns |
| Trépanier et al. (2007) | AFC | Gatineau (Canada) | Algorithms for inferring destination | A model estimating the destination for individual boarding buses with tap-in only data; 66% of destinations are well estimated |

Table 2 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-----------------------------|-----------------|---------------------------------|---|--|
| Zhao et al. (2007) | AFC & AVL & GIS | Chicago (USA) | Algorithms for identifying bus stop ID, bus ID, their integration, and OD inference | A method for inferring rail passenger trip OD matrices from an origin-only AFC system. A software tool is developed, and the results report the contributions of each inference method to OD estimations |
| Rossetti and Turitto (2000) | RFID signals | Virginia (USA) | Algorithms for the translation, storage, and retrieval of transit userdata | A prototype describing the integration of automatic passenger counting with automatic vehicle location to enable the automatic analysis of transit system performance |

A representative, but not a comprehensive list of references

AFC Automatic Fare Collection, *ANN* Artificial neural network, *APC* Automatic Passenger Counting, *AVL* Automatic Vehicle Location, *GIS* Geographic Information Systems, *OD* Origin–Destination, *RFID* Radio Frequency Identification, *SQL* Structured Query Language

Table 3 Studies on Portable device-based OD matrices

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|------------------------|-------------------|---------------------------------|--|---|
| Jee et al. (2023) | Wi-Fi & AVL | Kyoto (Japan) | MLR and ANN | A method for estimating the current queue length of bus passengers waiting at a stop and the waiting time for passengers at that location. The results show that randomised MAC addresses, non-randomised MAC addresses, and received signal strengths are critical parameters for determining queue length |
| Nitti et al. (2020) | Wi-Fi & APC | Cagliari (Italy) | Algorithms for de-randomization of MAC address and passenger counting | An advanced automatic bus passenger counting system for continuous urban mobility analysis, without requiring passenger involvement. The system effectively identifies the number of devices with active Wi-Fi interfaces: the accuracy rate is 100% in static scenarios and 94% in dynamic situations |
| Tu et al. (2019) | Wi-Fi & AFC & GIS | Shenzhen (China) | Algorithms for matching devices to users, trip labeling, feature extractions, network events | Investigating human mobility through the vehicular Internet service on public bus systems by mobile routers distributed in a city. The system can enhance the ability to observe passengers and their trips: smartcard observations are improved by 53.9% and 48.1%, respectively |
| Håkegård et al. (2018) | Wi-Fi & APC | Stavanger (Norway), | Algorithms for pre-processing of frames with local and global MAC address, and probability density functions | Applying statistical models for estimating OD matrices at the trip level. These models look more precise w.r.t. the cases in which one employs rigid thresholds for the median received signal level and observation time interval |

Table 3 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-------------------------|---------------------|---------------------------------|--|--|
| Lu et al. (2017) | SACDR | Singapore | Classifiers for kinematic motion, basic activity, advanced activity, learning algorithm | A framework for urban transportation service analysis merging participatory smartphone sensor data with city-wide transportation-related transactional data. The results from the subway boarding analyzer demonstrate the ability to identify failed boarding events with a precision larger than 90% |
| Chaudhary et al. (2016) | SACDR & GPS | Chandigarh (India) | Mobile applications and algorithms for prediction | A user-participation-based system designed to gather data about the occupancy levels in public transportation systems. The results show an accuracy on the occupancy level estimation of about 92% |
| Demissie et al. (2016) | Call Details Record | Senegal | Algorithms for identifying origins and destinations and their connections to trips taken from a transit network | A method for estimating public transport service passenger demand by cell phone data. The results showed prominent origins and destinations that are used to build OD matrices |
| Mishalani et al. (2016) | Wi-Fi & APC | Columbus (USA) | Algorithms for filtering out stationary devices, signals not considered on the bus and origin and destination stops of identified passengers | Comparing OD movements from Wi-Fi signal data with a combination of Wi-Fi and APC data with high-fidelity ground truth data obtained from an on-board survey. The results look better when Wi-Fi data are aggregated across multiple bus trips |

Table 3 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|-------------------------|---|------------------------------------|--|--|
| Carrel et al. (2015) | SACDR & AVL | San Francisco's Muni network (USA) | Algorithms for matching and inference of access and egress and transfer stops, and breakdowns to derive the final transit travel diary | A system for reconstructing and identifying the usage of transit by distinguishing between out-of-vehicle and in-vehicle portions of passenger trips. The results on the detection accuracy show a rate of about 93% |
| Deville et al. (2014) | Call Details Record | Portugal and France | Spatial statistics (Voronoi polygon) and power statistical models | Examining how aggregated CDR data might be used efficiently to map population distributions at the country scale. The method shows other unmeasurable patterns in space and time, while preserving the anonymity |
| Iqbal et al. (2014) | Call Details Record | Dhaka, Bangladesh | Algorithms for generating a tower-to-tower transient OD (t-OD) matrix, converting tower-to-tower t-OD to node-to-node t-OD, scaling and determining the actual OD matrix | A methodology developing OD matrix by mobile phone CDR and limited traffic count data. The results show the extracted trippatterns, while the counts are adopted to infer the ground truth traffic scenario |
| Calabrese et al. (2011) | Call Details Record census survey data | Metropolitan area of Boston, USA | Algorithms for trip determination and OD estimation | A method for the estimation of dynamic OD matrices through collected mobile phone location data. These estimations are well correlated with the US Census estimates |

Table 3 (continued)

| Authors (year) | Type of data | Location of data (city/country) | Methodology | Quick summary of contents and conclusions |
|------------------------|---------------------|---------------------------------|---|--|
| Caceres et al. (2008) | Call Details Record | Huelva and Seville (Spain) | Algorithms for location updating and paging | A method to correlate the flow of vehicles in a transportation network by leveraging on the flow of mobile phones in a cell-phone GSM network. The results are more cost-effective than those produced by conventional methods |
| Gonzalez et al. (2008) | Call Details Record | N/A | Some statistical models | Examining the paths of a substantial number of anonymized users tracked by mobile phones. Their paths look highly regular in space and time |

A representative, but not a comprehensive list of references

APC Automatic Passenger Counting, *AVL* Automatic Vehicle Location, *GIS* Geographic Information Systems, *GSM* global system for mobile communications, *OD* Origin–Destination, *SACDR* Smartphone app call details records, *CDR* Call details record

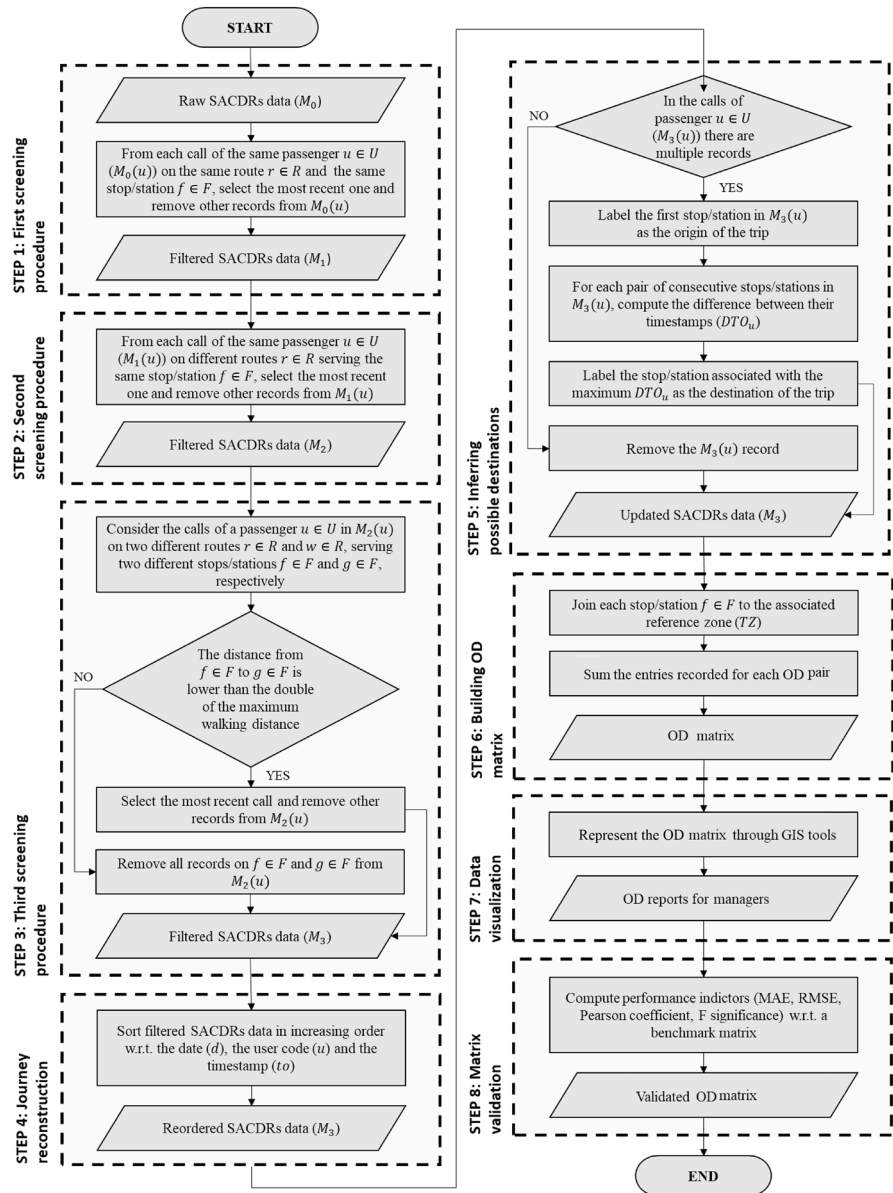


Fig. 1 Flowchart of the proposed methodology

3.1 Data type, preparation, and screening (STEPS from 1 to 3)

Consider a set of archived SACDRs, i.e., a database containing app call records (or queries) of passengers on routes and stops/stations of the network in the considered reference period (e.g., month, week). Although several app architectures

may exist with specific attributes, in this study the relevant attributes of each record are the call (query) code, date, route code, stop/station code, stop/station coordinates, user code (i.e., alphanumeric, and anonymous code generated when the passenger downloads the app) and the timestamp (i.e., the time when the passenger queries the app).

The first three steps pre-process the app data to reconstruct the journey of the passenger. Since all passengers can consult the app several times before their trip, one should remove redundant calls to bus stops and routes and consider a unique initial origin for their trip in a day. This is done by three data screening algorithms (procedures) to:

1. Remove redundant user calls on the same route and at the same bus stop/station (STEP 1).
2. Remove redundant user calls at the same bus stop/station of multiple routes, which are useful to join the destination (STEP 2).
3. Remove redundant user calls on different routes serving different stops/stations, which are useful to start a trip toward the destination (STEP 3).

These algorithms will be presented according to this notation. Let:

- U be the set of passengers using the app, R the set of routes and F the set of stops/stations.
- M_0 be the set of SACDRs in the transit network and $M_0(u)$ the subset of SACDRs generated by passenger $u \in U$.
- M_1 be the subset of records of M_0 returned at the end of the first screening algorithm (STEP 1) and $M_1(u)$ the subset of SACDRs of passenger $u \in U$ in this stage.
- M_2 be the subset of records of M_1 returned at the end of the second screening algorithm (STEP 2) and $M_2(u)$ the subset of SACDRs of passenger $u \in U$ in this stage.
- M_3 be the subset of records of M_2 returned at the end of the third screening algorithm (STEP 3) and $M_3(u)$ the subset of SACDRs of passenger $u \in U$ in this stage.

The relevant attributes of the i -th SACDR are: call code ic_i , date d_i , route code r_i , bus stop code f_i , coordinates X_{f_i} and Y_{f_i} , user code u_i , timestamp (or time occurrence) to_i .

3.1.1 First data screening procedure (STEP 1)

Sort M_0 in increasing order according to d_i , u_i and to_i . Assume that passenger $u \in U$ boards route $r \in R$ at stop/station $f \in F$. All passengers can query the app several times on stop/station $f \in F$ and route $r \in R$ to know the expected (real) vehicle arrival time, to time their arrival at $f \in F$ a few moments before the vehicle arrives.

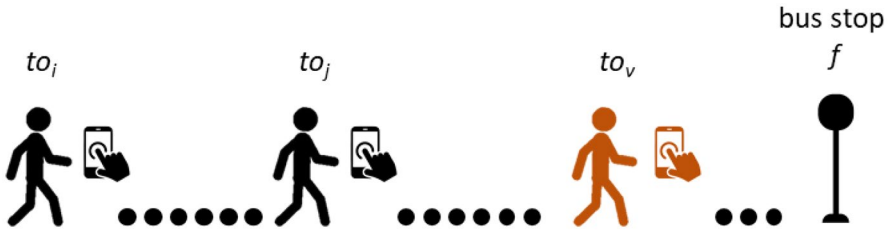


Fig. 2 Example of multiple calls to the smartphone app. The call selected is shown in orange

All queries of each passenger are recorded in the SACDRs database. Since passenger $u \in U$ can board only one route $r \in R$ at one stop/station $f \in F$, only one of their queries must be selected for inferring the origin of the trip.

The first screening algorithm selects the most recent timestamp to_i of passenger $u \in U$ to remove ambiguity among all their calls on route $r \in R$ at the same stop/station $f \in F$. More formally, for each passenger $u \in U$ querying for v times bus stop $f \in F$ and route $r \in R$, the algorithm selects the j -th record from $M_0(u)$ such that:

$$to_j = \max (to_1, \dots, to_v) \quad (1)$$

Next, the $v-1$ disregarded records are removed from $M_0(u)$ and the new list $M_1(u)$ is derived.

For instance, as shown in Fig. 2, passenger $u \in U$ queries stop/station $f \in F$ of route $r \in R$ on day d at timestamps $to_1 < to_j < to_v$. Three calls are included in the database, but only to_v is selected according to (1).

3.1.2 Second data screening procedure (STEP 2)

$M_1(u)$ lists the records of app calls at each stop/station $f \in F$ and/or route $r \in R$ for each passenger $u \in U$. Two or more routes may arrive at stop/station $f \in F$ to reach the same destination. Therefore, passenger $u \in U$ may query the app more than once on these different routes. Because passenger $u \in U$ can only board on a route for a trip, the second screening algorithm helps remove the ambiguity among different routes serving stop/station $f \in F$.

The actual route boarded cannot be known because the app does not provide this information. However, this is not a drawback because we aim to estimate the final OD matrix at the network level. The second screening algorithm selects the most recent timestamp among the different routes at the same stop/station $f \in F$. More formally, for each passenger $u \in U$ who queries all useful routes $r \in R$ at stop/station $f \in F$ for q times, this procedure selects the j -th record in $M_1(u)$ such that:

$$to_j = \max (to_1, \dots, to_q) \quad (2)$$

Next, the $q - 1$ disregarded records are removed from $M_1(u)$ and the new list $M_2(u)$ is derived.

3.1.3 Third data screening (STEP 3)

$M_2(u)$ lists records that contain single app calls at each stop/station $f \in F$ and route $r \in R$ for passenger $u \in U$. However, there may be more than one pair of buses and routes which are useful to join the destination. For example, one could take route $r \in R$ at stop/station $f \in F$ or route $w \in R$ at stop/station $g \in F$. Therefore, each passenger $u \in U$ may consult the app by querying the expected (real) bus arrival times of route $r \in R$ at stop/station $f \in F$ and route $w \in R$ at stop/station $g \in F$, respectively, to select (and reach) the most convenient path for their trip.

This situation results in two or more calls to the app by passenger $u \in U$. Since they can board only one route at one stop/station, the third screening algorithm removes the ambiguity among these calls. It works as follows.

If one knows the location of the passenger querying the app, it can be adopted to estimate the distance from the bus stops of interest. If the maximum acceptable walking distance is δ , one could assume that the passenger opts for the closest bus stop. However, this assumption does not hold in this framework, because passengers are not tracked by the app. The only known distances are those between queried bus stops. If these distances are beyond a threshold depending on δ , the queries of the customer to these bus stops are removed, else the queries of the passenger are processed, and the most recent query identifies the selected bus stop. A reasonable value of the threshold can be computed in the case of passengers located at distance δ from both bus stops $f \in F$ and $g \in F$, i.e., the threshold can be set to 2δ .

More formally, the distance $dist_{fg}$ between stops/stations f and g is taken. If $dist_{fg}$ is larger than 2δ , the queries of user $u \in U$ are disregarded, else all q time occurrences of user $u \in U$ at bus stops $f \in F$ and $g \in F$ are taken, sorted in ascending order and the last value (i.e., the maximum) is selected to define the departure bus stop. If to^f denotes the latest time occurrence associated with the query at bus stop f , this procedure can be described as follows for each user u :

$$\left\{ \begin{array}{l} \text{if } (0 \leq dist_{fg} \leq 2\delta) \\ \quad \{ \text{if } to^f = \max(to_1, \dots, to_z) \\ \quad \quad \text{return the record containing } f \\ \quad \text{else} \\ \quad \quad \text{return the record containing } g \\ \quad \quad \text{else} \\ \quad \quad \text{remove the records containing } f \text{ and } g \text{ from the } M_2(u) \end{array} \right. \quad (3)$$

Equation (3) selects the last queried stop/station and route, which may not be used by passenger u because these data cannot be gathered from the app. However, this is not a drawback because the final OD matrix will be obtained at the network level, and both stops/stations f and g are close to each other. Therefore, either bus stop $f \in F$ or bus stop $g \in F$ may approximate the origin of the trip.

Next, the $z - 1$ records disregarded are removed from $M_2(u)$ and the new list $M_3(u)$ is derived.

Notably, even if the former description focused only on two routes and two stops/stations, the third data screening algorithm can be generalised for multiple routes

and stations. In this case, because passengers are supposed to board only one route at a time, it is sufficient to take all pairs of stops/stations (and routes) of interest and repeat the previous screening.

The pseudo code of the **Data screening** algorithm is reported hereafter:

Algorithm 1 Data Screening // Screen SACDR to remove anomalies in data

```

For each  $i \in M_0(u)$  do // first data screening
  if ( $d_i = d_{i+1}$  and  $r_i = r_{i+1}$  and  $f_i = f_{i+1}$ ) and ( $to_{i+1} > to_i$ )
    return  $M_1(u) = M_0(u) - \{i\}$  // return  $M_0(u)$  without record  $i$ 
  else
    return  $M_1(u) = M_0(u) - \{i+1\}$  // return  $M_0(u)$  without record  $i+1$ 
  end if
end for
update  $M_1(u)$  // return  $M_1(u) \subseteq M_0(u)$ 

For each  $i \in M_1(u)$  do // second data screening
  if ( $d_i = d_{i+1}$  and  $f_i = f_{i+1}$ ) and  $r_i \neq r_{i+1}$  and ( $to_{i+1} > to_i$ )
    return  $M_2(u) = M_1(u) - \{i\}$  // return  $M_1(u)$  without record  $i$ 
  else
    return  $M_2(u) = M_1(u) - \{i+1\}$  // return  $M_1(u)$  without record  $i+1$ 
  end if
end for
update  $M_2(u)$  // return  $M_2(u) \subseteq M_1(u)$ 

For each  $i \in M_2(u)$  do // third data screening
  if ( $d_i = d_{i+1}$ ) and ( $f_i \neq f_{i+1}$  and  $r_i \neq r_{i+1}$ )
    if  $0 \leq \text{dist}_{f_i f_{i+1}} \leq 2\delta$ 
      if  $to^{f_i} = (to_1, \dots, to_q)$ 
        return  $M_3(u) = M_2(u) - \{i+1\}$  // return  $M_2(u)$  without record  $i+1$ 
      else
        return  $M_3(u) = M_2(u) - \{i\}$  // return  $M_2(u)$  without record  $i$ 
      end if
    else
      return  $M_3(u) = M_2(u) - \{i\} - \{i+1\}$  // return  $M_2(u)$  without records  $i$  and  $i+1$ 
    end if
  end if
end for
update  $M_3(u)$  // return  $M_3(u) \subseteq M_2(u)$ 

```

3.2 Journey reconstruction (STEP 4)

At the end of the data preparation and screening stage, $M_3(u)$ lists records that contain single app calls at each bus stop $f \in F$, route $r \in R$ and passenger $u \in U$. $M_3(u)$ is sorted in increasing order w.r.t. d , u and to .

Since each record represents a call made before boarding a bus and a journey could be rebuilt from the sequence of calls, a new attribute can be added to indicate the progressive stop/station number queried by passenger $u \in U$. It is denoted by PO_i , if it is associated with the i -th call of passenger u and let $PO(u)$ be a vector with entries PO_i . Therefore, each passenger $u \in U$ can be associated with a journey consisting of a sequence of the called stops/stations, where each segment of the journey starts.

The pseudo code of the **Journey reconstruction** algorithm is as follows:

Algorithm 2 Journey reconstruction // sort for the times of occurrence to rebuild the journey of a passenger

```

For each  $i \in M_3(u)$  do // examine each call done by each user
  Sort ( $to_i$ )
  Initialize  $PO_i = 1$ 
  While  $\{(PO_i \leq |M_3(u)|)\}$ 
    Increase  $PO_{i+1} = PO_i + 1$ 
  End while
end for
update  $M_3(u) = (M_3(u) \cup \{PO(u)\})$  // the new attribute  $PO(u)$  is added to  $M_3$ 

```

3.3 Inferring possible destinations (STEP 5)

The records in $M_3(u)$ represent the pre-processed calls of passenger $u \in U$ before boarding a bus. At this stage, they define the sequence of bus stops boarded in the passenger's journey. Since there is no app call on alighting bus stops, the destination bus stops (through possible transfers) are not known and must be estimated for each passenger $u \in U$ from their journey $M_3(u)$. The following procedure is proposed for this estimation:

- (i) If there is only a record in $M_3(u)$, we cannot determine an alighting location, because passenger $u \in U$ could alight at each bus stop.
- (ii) If there are multiple records in $M_3(u)$, a destination bus stop can be detected among those after the first one.

In case (i), the record is disregarded. Because we are interested in building a daily OD matrix for public transportation, this choice is not a strong drawback, because at the passenger's return home using some other ways (e.g., by car or bus when passengers and vehicles arrive simultaneously, thus avoiding passenger u having to query the app).

In case (ii), the first bus stop in $M_3(u)$ is labelled as the origin of the trip. Next, let h be the index of stop/station after the first one in $M_3(u)$. For each pair of consecutive stops/stations, the difference DTO_u between their timestamps is computed for each passenger $u \in U$ and the maximum value is selected:

$$DTO_u = \max_h (to_h - to_{h-1}) \quad (4)$$

The bus stop associated with this maximum value is taken as the alighting location and is supposed to be the destination.

Figure 3 shows the possible destinations among bus stops f_1, f_2, f_3 and f_4 for passenger u . In Fig. 3, passenger $u \in U$ queries the app from their origin by imputing the boarding route r at bus stop f_1 in the origin TAZ at timestamps to_1 . Moreover, passenger $u \in U$ queries the app at boarding bus stops f_2 at to_2 , f_3 at to_3 and f_4 at to_4

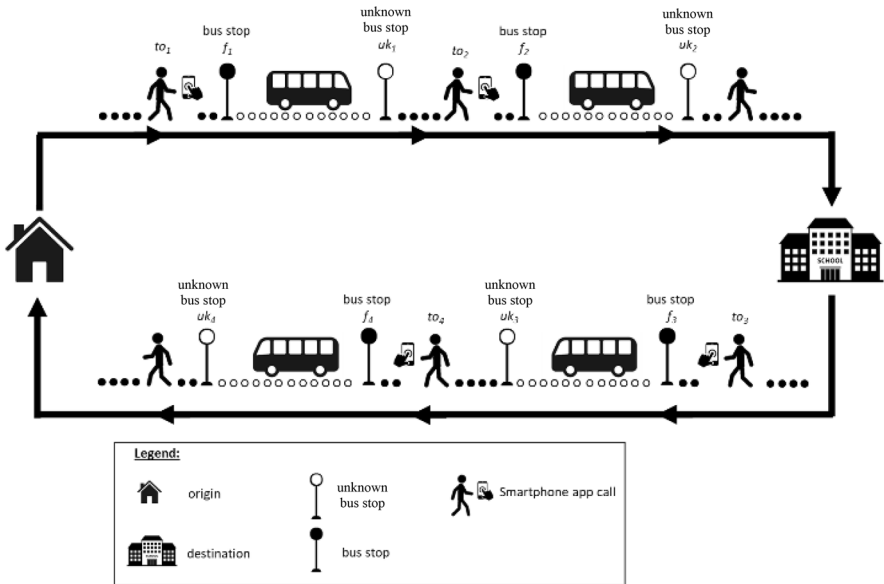


Fig. 3 Example of travel chain from the origin to the destination

such that $to_1 < to_2 < to_3 < to_4$, but one does not know the alighting bus-stops a_1, a_2, a_3 and a_4 of these four trip segments. According to Eq. (4),

$$DTO_u = \max (to_2 - to_1; to_3 - to_2; to_4 - to_3) \tag{5}$$

If $DTO_u = to_3 - to_2$, f_3 is set to be the departure bus stop of the reverse trip segment from the destination to the origin. Moreover, f_3 is set to be the alighting bus stop of the former trip segment and belongs to the destination TAZ (i.e., $f_3 = a_2$). Moreover, we assume that passengers end the last trip of that day at the stop where they boarded at the beginning of the journey (i.e., $f_1 = a_4$).

The pseudo code of the **Inferring the destination** algorithm is as follows:

Algorithm 3 Inferring the destination// Inferring origin and destination

```

Define  $O_u$ =origin bus stop among the  $i$  calls in  $M_3(u)$ ,  $D_u$ =destination bus stop among the  $i$  calls in  $M_3(u)$ 
For each  $i \in M_3(u)$  do
  if  $\max \{PO_i\} = 1$  // if there is 1 record only for passenger  $u$ 
    return  $M_3(u) = M_3(u) - \{i\}$  // remove record  $i$  from  $M_3(u)$ 
  else
     $O_u = \{f \in F : PO_i = 1\}$  // selection of the departure bus stop
    While  $\{PO_i \leq |M_3(u)|\}$ 
       $DTO = \max (to_i - to_{i-1})$  // compute the maximum value between  $to$  and  $to-1$ 
       $D_u = \{f \in F : PO_i = PO_i(DTO)\}$  // selection of the destination bus stop
    end while
  end if
end for
update  $M_3(u)$ 

```

3.4 Building OD matrix (STEP 6)

Because we are interested in building the OD matrix of the transit network, an existing zoning of the investigated context should be considered. Specifically, a special area consisting of one or more census blocks, block groups, or census tracts was delineated by the state and/or local transportation officials. This is used to tabulate traffic-related data, such as home-to-work and home-to-study trip statistics. Therefore, matching each bus stop to each TAZ before inferring the matrix is required. Specifically, the procedure is as follows. Let:

- TZ be the set of TAZs and $c \in TZ$ the centroid of TAZ, which has coordinates X_c and Y_c .

Next, compute the distance $dist_{cf}$ between each centroid $c \in TZ$ and bus stop $f \in F$. If this distance is lower than a threshold δ , the stop station could belong to the TAZ. In order to remove the ambiguity of a stop station in multiple TAZs, only the minimum distance w.r.t. all centroids is considered for a bus stop, because all passengers are supposed to board/alight at the stop closest to their origin/destination.

Next, a new attribute is added to $M_3(u)$ to describe the centroid (and the associate TAZ) of each origin bus stop O_u or destination bus stop D_u of passenger $u \in U$. This attribute is denoted by c_f^* for bus stop $f \in F$. Once each bus stop is joined to the reference zone, the final matrix is built by summing the entries recorded for each OD pair.

The pseudo code of the **Building OD matrix** algorithm is reported hereafter:

Algorithm 4 Building OD Matrix // stops/stations assigned to a specific TAZ, M_3 is enhanced with new attribute called TAZ code, sum for each OD pair with the same TAZ.

```

 $C_f = \emptyset$  // set of possible centroids of bus stop  $f \in F$ 
 $dist_{cf}$  = distance between centroid  $c$  and bus stop  $f \in F$ 
For each  $u \in U$ 
  For each  $c \in TZ$ 
    If  $f=O(u)$  or  $f=D(u)$ 
      If  $0 \leq dist_{cf} \leq \delta$ 
         $C_f = \{C_f \cup c\}$ 
      end if
     $c_f^* = \{c \in C_f: dist_{cf} = \min\{dist_{cf}\}\}$  // centroid associated with bus stop
    update  $M_3(u) = (M_3(u) | c_f^*)$  // the new attribute  $PO(u)$  is added to  $M_3(u)$ 
  end if
end for
end for

O=set of origin TAZs; D=set of TAZs destination;
For each  $o \in O$ 
  For each  $d \in D$ 
    Count_od=0
    For each  $u \in U$ 
      If  $c_{O(u)}^* \in O$  and  $c_{D(u)}^* \in D$  //  $c_{O(u)}^*$  and  $c_{D(u)}^*$  are the centroids of the origin and the
        // destination of passenger  $u \in U$ 
        Count_od = Count_od + 1;
      end if
    end for
  end for
end for
end for

```

Although this matrix was built by data collected on a daily time scale, the overall procedure can be run in shorter periods of interest.

3.5 Data visualisation (STEP 7)

Building comprehensible and usable OD reports is a key factor for the effective analysis of SACDRs. Therefore, it is relevant to study and develop appropriate routines to generate brief and relevant reports that are easily understandable for planners, senior managers, and decision makers. In this context, clear representations of the OD matrix SACDRs can be achieved using GIS tools such as Qgis (see <https://www.qgis.org/en/site/>) to highlight which TAZs deserve more attention and could benefit from the improvement in the service.

3.6 Matrix validation (STEP 8)

The performance of the framework is analysed by comparing the resulting OD matrix against a suitable benchmark matrix. The comparison is done in terms of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), to measure the average magnitude of the errors in a set of forecasts, without considering their direction. Both MAE and RMSE are negatively oriented scores (i.e., the lower the values, the better the results). More formally, if we denote by $SACDR_{o,d}$ and $STAT_{o,d}$ the forecast and the benchmark data for the generic origin $o \in O \subseteq TZ$ and the generic destination $d \in D \subseteq TZ$, respectively, these statistical indicators are computed for each origin as follows:

$$MAE_o = \frac{1}{|D|} \sum_{d \in D} |STAT_{o,d} - SACDR_{o,d}| \quad \forall o \in O \subseteq TZ \quad (6)$$

$$RMSE_o = \sqrt{\frac{\sum_{d \in D} |STAT_{o,d} - SACDR_{o,d}|^2}{|D|}} \quad \forall o \in O \subseteq TZ \quad (7)$$

Similarly, as for the generic destination $d \in D \subseteq TZ$, the error indicators are computed according to the Eqs. (6) and (7), where o is replaced with d .

Additionally, the Pearson correlation coefficient, denoted by $r_{ISTAT,SACDR}$, is also computed. It is a positively oriented score, where a value closer to unity indicates a better result. More formally, as for the generic origin $o \in O \subseteq TZ$, the Pearson correlation coefficient is computed as indicated in Eq. (8).

$$r_{ISTAT,SACDR_o} = \frac{\sum_{d \in D} [(STAT_{o,d} - \underline{STAT}_o)(SACDR_o - \underline{SACDR}_o)]}{\sqrt{\sum_{d \in D} [(STAT_{o,d} - \underline{STAT}_o)^2 (SACDR_o - \underline{SACDR}_o)^2]}} \quad \forall o \in O \subseteq TZ \quad (8)$$

Similarly, as for the generic destination $d \in D \subseteq TZ$, the Pearson correlation coefficient is computed according to the Eq. (8), where o is replaced with d .

Furthermore, for each generic origin $o \in O \subseteq TZ$, a Linear Regression Model (LRM) is built to evaluate the link between the target number of trips (i.e., $STAT_{o,d}$) and the predicted one (i.e., $SACDR_{o,d}$):

$$SACDR_{o,d} = \alpha_o + \beta_o \cdot STAT_{o,d} \quad \forall o \in O \subseteq TZ \quad (9)$$

where α_o and β_o are the regression coefficients. Similarly, a regression model is built for each destination $d \in D \subseteq TZ$ (in Eq. (9), index o is replaced with d). The statistical significance of each LRM (denoted as *p-value*) is determined by performing an F-test.

4 Real-case experiment

4.1 Context

The overall framework has been tested in the area of Cagliari, which has about 400,000 inhabitants and is located on the island of Sardinia (Italy). CTM is the local public transport company and oversees public transportation using 271 vehicles (buses and trolleys). Moreover, these vehicles travel over 12.4 million kilometres annually along 34 routes and serve over 90% of the transit passenger demand in this area (CTM 2020).

The e-age and recent developments in portable devices have pushed CTM to develop an app, “BusFinder”,² for mobile devices such as smartphones and tablets. It provides passengers with all-in-one pre-trip and en-route real-time information of routes and bus stops in the area of Cagliari (Tilocca et al. 2017). Specifically, passengers can plan their trips in different ways: (i) inputting the origin and destination addresses by a short text, (ii) selecting the origin and destination by a pointer on a map, and (iii) using the GPS coordinates of their smartphone like a navigator tool. Moreover, by selecting a bus stop of interest, the passenger receives information on useful routes and the expected bus arrival time for the next bus arrival (Fig. 4). The left side of Fig. 4 shows general information (i.e., on bus stop ‘Dante (ang. via Todde)’ and two different routes are available, #1 and #3). The central part of this figure lists some bus stops of interest for a passenger and the pedestrian distance to reach them. The right side of this figure

² In Google Play the following statements are reported about the app “BusFinder”. “Data that may be shared with other companies or organisations. Data practices may vary based on your app version, use, region, and age”. To clarify, after the download, an anonymised user code is automatically generated and kept until the uninstallation. The app can be used in three ways. The first way enables passengers to buy a travel ticket after a registration by a user code and a password. The passenger must agree with the privacy policy of CTM and is informed about the processing of personal data. The second way enables users to plan a trip along the overall PT network and the app asks permission to track passengers at the origin. They should provide the consent by a function activating the localisation in the map and informing on the processing of personal data. The third way enables passengers to query the app on travel information about the real arrival times in a specific bus stop without providing any consent, because passengers are not tracked. This way is used in this research.

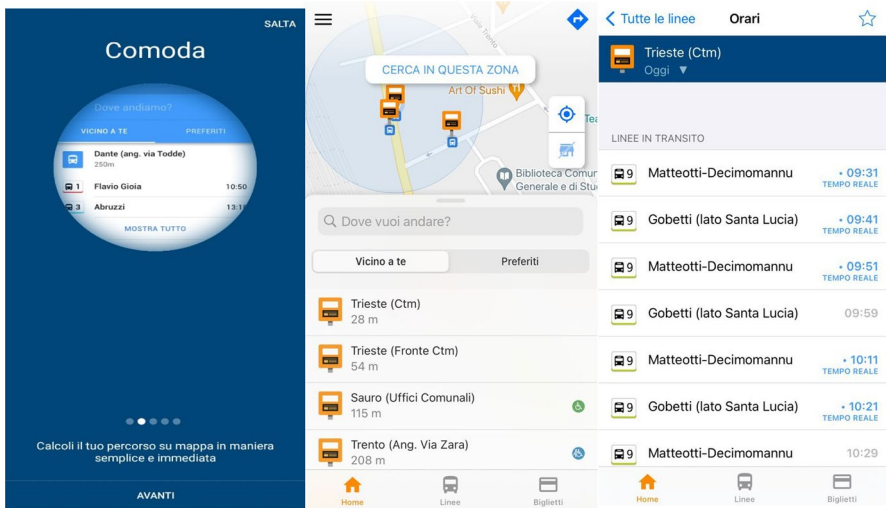


Fig. 4 Overview of “BusFinder”

reports an overview of bus arrival times (in blue). For instance, the first vehicle for route 9—direction ‘Matteotti–Decimomannu’—at bus stop ‘Trieste CTM’ is expected to arrive at 09:31.

These features were achieved by integrating archived information on routes and sequences of bus stops gathered in a GIS environment, and real-time information on expected bus arrival (at selected bus stops) provided by AVL technology, all of which have been equipped on buses since 2007.

According to the last available statistics, CTM is quite satisfied with this app because it has been downloaded by several stores (e.g., IOS, Android, and Windows Phone). Since its launch, “BusFinder” was downloaded approximately 180,000 times. It is adopted by more than 80% of CTM’s passengers (CTM 2020).

4.2 Experimental setup and results

The method was developed and implemented on MS Access and MS Excel, as well as Qgis running on a standard PC. In the experimentation, we gathered 9.6 M + raw SACDRs data collected daily from 01/01/2021 to 31/12/2021. These correspond to approximately 0.8 M SACDRs data monthly.

Next, drawing on these data, 30 workdays were selected from the middle of September to the end of November. Hence, about 1.5 M + SACDRs were considered to show the viability of the method in Sect. 3. In what follows, the execution of the method is described on a specific day and the average values of MAE, RMSE and r are computed over the 30 workdays for both origins and destinations.

Figure 5 provides a portion of the original M_0 on a specific day, according to Sect. 3. This shows that user *GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3et-dAheYY936kihTg* queried “BusFinder” 12 times on 21/10/2021. Specifically, the

| call code | bus stop code | route code | user code | day | time occurrence |
|-----------|---------------|------------|--|------------|-----------------|
| 4855 | IT0549 | 8A | GsDId-002mExxFQQuaxuOkY4b0eSlxUjIDKFOA4H8H9kgLDHnM | 21/10/2021 | 09:47:44 |
| 33847 | CF0043 | 1 | GsDId-00lgwwQNAO2QLI87LPDTCGcC25lNJUcTkuGnHtzp0X | 21/10/2021 | 06:52:54 |
| 21753 | HB0520 | 1 | GsDId-00Cs5FdgnF8re0IrtzfwaklXhTnIzVut1IDBHQINi2O | 21/10/2021 | 18:21:17 |
| 10438 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:56:00 |
| 10521 | AB110 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:57:21 |
| 10676 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:59:06 |
| 10688 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:59:40 |
| 10992 | PM0030 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 08:19:45 |
| 11120 | PM0030 | 30 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 08:20:08 |
| 17658 | SO0020 | 30 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:39:00 |
| 18105 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:40:31 |
| 18182 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:44:09 |
| 19128 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:45:12 |
| 20775 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:56:23 |
| 21423 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:57:07 |
| 30024 | DP0989 | 40 | GsDId-01Fy4Xwyk9f00ThftDGKfFDGryIXZJRoppFshArtma1J | 21/10/2021 | 22:08:55 |
| 30025 | DP1102 | 41 | GsDId-01Fy4Xwyk9f00ThftDGKfFDGryIXZJRoppFshArtma1J | 21/10/2021 | 22:07:59 |
| 41669 | DP1102 | 41 | GsDId-01Fy4Xwyk9f00ThftDGKfFDGryIXZJRoppFshArtma1J | 21/10/2021 | 07:54:39 |
| 46710 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7PhI9Rv0rFgx2TXREwXucy | 21/10/2021 | 08:37:39 |
| 59461 | CU0741 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7PhI9Rv0rFgx2TXREwXucy | 21/10/2021 | 12:59:56 |
| 59485 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7PhI9Rv0rFgx2TXREwXucy | 21/10/2021 | 12:59:31 |
| 48834 | PG0221 | 29 | GsDId-01NgsgzlkNSbcPGYzsr46Dgc3VPNAZFZP0g3XKHHzP | 21/10/2021 | 09:04:04 |
| 39617 | LB0315 | 3 | GsDId-01W1fCakYjnahTx8wC1ItAxQM85jIteVwVhkihszVt4u | 21/10/2021 | 07:32:48 |
| 40635 | TR0311 | 3 | GsDId-01W1fCakYjnahTx8wC1ItAxQM85jIteVwVhkihszVt4u | 21/10/2021 | 07:33:03 |
| 28921 | NA0620 | 9 | GsDId-024ff6ScnkNQCuZxcrcClk55ER2LDTYshW8nonw62kWMME | 21/10/2021 | 21:18:37 |
| 33099 | AF1176 | 9 | GsDId-026038alUxaE2376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:37:55 |
| 33689 | AF1176 | 9 | GsDId-026038alUxaE2376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:53:53 |
| 34611 | AF1176 | 9 | GsDId-026038alUxaE2376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:02:26 |
| 39914 | NS0158 | 5 | GsDId-026038alUxaE2376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:44:51 |
| 40294 | RM2109 | 5 | GsDId-026038alUxaE2376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:45 |
| 40638 | TS0161 | 5 | GsDId-026038alUxaE2376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:48 |
| 20976 | DD0766 | QS | GsDId-026SCYyGT45HbU0LDz652x73KhqBr0tG2L0G4GpXQA9 | 21/10/2021 | 18:12:39 |
| 21552 | BR0503 | QS | GsDId-026SCYyGT45HbU0LDz652x73KhqBr0tG2L0G4GpXQA9 | 21/10/2021 | 18:15:28 |
| 22488 | NO0701 | PQ | GsDId-026SCYyGT45HbU0LDz652x73KhqBr0tG2L0G4GpXQA9 | 21/10/2021 | 18:38:02 |

Fig. 5 A portion of M_0 : Raw data sample gathered from the SACDRs database

passenger queried routes 1 and 31 four times, route 9 thrice, and route 30 once at different times (see the red-highlighted section in Fig. 5).

Next, according to STEP 1, the first data screening algorithm was applied: all SACDRs with the same passenger $u \in U$, route $r \in R$, and boarding stop/station $f \in F$ to the origin were removed. The most recent timestamp was kept according to Eq. (1).

Figure 6 shows a simple example on user $GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg$: the yellow-highlighted records were removed. In three records related to route 9, only the last is retained because it has the maximum to . Thus, five records were deleted for this user. At the end of this screening procedure, 32,906 records were available (on average over the 30 workdays), and M_1 was returned.

Next, according to STEP 2, the second data screening algorithm was applied: all SACDRs with the same passenger $u \in U$ and the same bus stop served by two or more routes were removed, except those with the most recent timestamp to according to Eq. (2). Figure 7 shows a simple example for user $GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg$: the violet-highlighted records were removed. Specifically, bus stop AB0110, which is common for routes 1 and 9, was considered. The violet-highlighted record related to route 1 is removed because its timestamp to is earlier than that of route 9; thus, the removed record refers to a previous call at that bus stop. Hence, only the record related to route 9 is retained. To summarise,

| call code | bus stop code | route code | user code | day | time occurrence |
|-----------|---------------|------------|---|------------|-----------------|
| 4855 | IT0549 | 8A | GsDId-002mExxFfQUaxuOkY4b0eSlxUjIDKFOA4H89kgLDhNm | 21/10/2021 | 09:47:44 |
| 33847 | CF0043 | 1 | GsDId-00lgrwvQNAO2QLI87LPDTCGc25lNJUCtku6nHtzp0XT | 21/10/2021 | 06:52:54 |
| 21753 | H80520 | 1 | GsDId-00oSfDgnF8re0lrtzfwaklXhTnlZiVut1IDBHQINi2O | 21/10/2021 | 18:21:17 |
| 10438 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:56:00 |
| 10521 | AB110 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:57:21 |
| 10676 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:59:06 |
| 10688 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:59:40 |
| 10992 | PM0030 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 08:19:45 |
| 11120 | PM0030 | 30 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 08:20:08 |
| 17658 | S00020 | 30 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:39:00 |
| 18105 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:40:31 |
| 18182 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:44:09 |
| 19128 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:45:12 |
| 20775 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:56:23 |
| 21423 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:57:07 |
| 30024 | DP0989 | 40 | GsDId-01Fy4Xwyk9fo0ThftDGKFDGryIXZJR0ppFsHaRtmaJJ | 21/10/2021 | 22:08:55 |
| 30025 | DP1102 | 41 | GsDId-01Fy4Xwyk9fo0ThftDGKFDGryIXZJR0ppFsHaRtmaJJ | 21/10/2021 | 22:07:59 |
| 41669 | DP1102 | 41 | GsDId-01Fy4Xwyk9fo0ThftDGKFDGryIXZJR0ppFsHaRtmaJJ | 21/10/2021 | 07:54:39 |
| 46710 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jQ3X1O7PhI9RV0fgx2TXREwXucY | 21/10/2021 | 08:37:39 |
| 59461 | CU0741 | 29 | GsDId-01Moy2EfpzLzHV2jQ3X1O7PhI9RV0fgx2TXREwXucY | 21/10/2021 | 12:59:56 |
| 59485 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jQ3X1O7PhI9RV0fgx2TXREwXucY | 21/10/2021 | 12:59:31 |
| 48834 | PG0221 | 29 | GsDId-01NgszlkNSbcPGYxz4r6Dgc3VPNAFZP0g3xKkHhZp | 21/10/2021 | 09:04:04 |
| 39617 | L80315 | 3 | GsDId-01Wl1fCaKjYnahTx8wC1ItAxQM85jIleWvHkhszVt4u | 21/10/2021 | 07:32:48 |
| 40635 | TR0311 | 3 | GsDId-01Wl1fCaKjYnahTx8wC1ItAxQM85jIleWvHkhszVt4u | 21/10/2021 | 07:33:03 |
| 28921 | NA0620 | 9 | GsDId-024ff6ScnkNQCuzxcrCik5SER2LDTYshW8nonw62kWMME | 21/10/2021 | 21:18:37 |
| 33099 | AF1176 | 9 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:37:55 |
| 33689 | AF1176 | 9 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:53:53 |
| 34611 | AF1176 | 9 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:02:26 |
| 39914 | NS0158 | 5 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:44:51 |
| 40294 | RM2109 | 5 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:45 |
| 40638 | TS0161 | 5 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:48 |
| 20976 | DD0766 | QS | GsDId-026SCYyGT45IHbU0LdZ652x73KhqBrOtG2L0G4GpXQA9 | 21/10/2021 | 18:12:39 |
| 21552 | BR0503 | QS | GsDId-026SCYyGT45IHbU0LdZ652x73KhqBrOtG2L0G4GpXQA9 | 21/10/2021 | 18:15:28 |
| 22488 | NO0701 | PQ | GsDId-026SCYyGT45IHbU0LdZ652x73KhqBrOtG2L0G4GpXQA9 | 21/10/2021 | 18:38:02 |

Fig. 6 A portion of M_1 : Raw data sample after the first screening algorithm (STEP 1)

two records were deleted for this user and five records were retained at the end of the second screening procedure. Finally, after applying the second screening procedure for each user, a total of 22,466 records were available on this day (on average over the 30 workdays) and M_2 was returned.

As illustrated in STEP 3, the third data screening algorithm was applied: all SACDRs with the same passenger $u \in U$ at stops/stations $f \in F$ and $g \in F$ served by routes $r \in R$ and $w \in R$, respectively, were removed except that with the most recent timestamp to , according to Eqs. (3).

Figure 8 shows a simple example. For ease, the same user $GsDId-04kcs-vU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg$ is considered.

For this experiment, δ was set to 400 m (e.g., Murray et al. 1998). For instance, the distance $dist_{fg}$ between bus stops S0020 and DA1268 was taken. Next, since $dist_{fg} \leq 2\delta$ and bus stop DA1268 has the largest to , it was selected. Therefore, the green-highlighted record of bus stop S0020 was removed. To summarise, one record was deleted for this user and four records were kept at the end of the third screening procedure, as shown in Fig. 9. Finally, after applying the third screening algorithm for each user, 14,501 records were available on this day and M_3 was returned.

Next, according to STEP 4, for each passenger u , we can build the daily journey that consists of the sequence of queried (and validated) stops/stations. These are arranged in ascending order according to the timestamps. As for user $GsDId-04kcs-vU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg$, the sequence of queried bus

| call code | bus stop code | route code | user code | day | time_occurrence |
|-----------|---------------|------------|---|------------|-----------------|
| 4855 | IT0549 | 8A | GsDId-002mExxFfQUaxuOkY4b0eSlxUIIDKfOa4H8H9kLDhNm | 21/10/2021 | 09:47:44 |
| 33847 | CF0043 | 1 | GsDId-00lgwwQNAO2QLI87LPDTCGc251NuUctku6nHtzp0XT | 21/10/2021 | 06:52:54 |
| 21753 | HB0520 | 1 | GsDId-00OcsSfdgnF8re0irtzfwaklXhTnlzVut11DBHQI20 | 21/10/2021 | 18:21:17 |
| 10438 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:56:00 |
| 10521 | AB110 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:57:21 |
| 10676 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:59:06 |
| 10688 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 07:59:40 |
| 10992 | PM0030 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 08:19:45 |
| 11120 | PM0030 | 30 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 08:20:08 |
| 17658 | SO0020 | 30 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:39:00 |
| 18105 | DA1268 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:40:31 |
| 18182 | DA1268 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:44:09 |
| 19128 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:45:12 |
| 20775 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:56:23 |
| 21423 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg | 21/10/2021 | 16:57:07 |
| 30024 | DP0989 | 40 | GsDId-01Fy4Xwyk9f0ThftDGKFFDGryiXZJR0ppFshRtma1J | 21/10/2021 | 22:08:55 |
| 30025 | DP1102 | 41 | GsDId-01Fy4Xwyk9f0ThftDGKFFDGryiXZJR0ppFshRtma1J | 21/10/2021 | 22:07:59 |
| 41669 | DP1102 | 41 | GsDId-01Fy4Xwyk9f0ThftDGKFFDGryiXZJR0ppFshRtma1J | 21/10/2021 | 07:54:39 |
| 46710 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7Ph9Rv0fGx2TXREuWxucY | 21/10/2021 | 08:37:39 |
| 59461 | CU0741 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7Ph9Rv0fGx2TXREuWxucY | 21/10/2021 | 12:59:56 |
| 59485 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7Ph9Rv0fGx2TXREuWxucY | 21/10/2021 | 12:59:31 |
| 48834 | PG0221 | 29 | GsDId-01NgsgzlkNSbcPGYzsz4r6Dgc3VPNAEFZP0gq3XKXhZP | 21/10/2021 | 09:04:04 |
| 39617 | LB0315 | 3 | GsDId-01W1fCakYjnahTx8wC1ItAxQM85jIleVvHkhszVt4u | 21/10/2021 | 07:32:48 |
| 40635 | TR0311 | 3 | GsDId-01W1fCakYjnahTx8wC1ItAxQM85jIleVvHkhszVt4u | 21/10/2021 | 07:33:03 |
| 28921 | NA0620 | 9 | GsDId-024ff6ScnkNQcuzxrcClik5SER2LDTYshW8nonw6ZKWME | 21/10/2021 | 21:18:37 |
| 33099 | AF1176 | 9 | GsDId-026O3BalUXaE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:37:55 |
| 33689 | AF1176 | 9 | GsDId-026O3BalUXaE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:53:53 |
| 34611 | AF1176 | 9 | GsDId-026O3BalUXaE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:02:26 |
| 39914 | NS0158 | 5 | GsDId-026O3BalUXaE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:44:51 |
| 40294 | RM2109 | 5 | GsDId-026O3BalUXaE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:45 |
| 40638 | TS0161 | 5 | GsDId-026O3BalUXaE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:48 |
| 20976 | DD0766 | QS | GsDId-026SCYyGT45IHbU0LDz652x73kHqBrOtG2L0G4GpXQA9 | 21/10/2021 | 18:12:39 |
| 21552 | BR0503 | QS | GsDId-026SCYyGT45IHbU0LDz652x73kHqBrOtG2L0G4GpXQA9 | 21/10/2021 | 18:15:28 |
| 22488 | NO0701 | PQ | GsDId-026SCYyGT45IHbU0LDz652x73kHqBrOtG2L0G4GpXQA9 | 21/10/2021 | 18:38:02 |
| 5094 | BD0848 | PF | GsDId-02HndsuldgKh8Og1UUA8TK1A3PrN1qNVDrD8f6obgre | 21/10/2021 | 10:01:01 |

Fig. 7 A portion of M_2 : Raw data sample after the second screening algorithm (STEP 2)

stops during a day is AB0110, PM0030, DA1268, and finally, RM0739. Because these queried bus stops are origin points, we need to infer the destination, according to the method described in STEP 5. Specifically, we obtain $M_3(u) = \{AB0110, PM0030, DA1268, RM0739\}$. Since the journey $M_3(u)$ contains multiple origins, we can find a destination because the passengers alighted at some stops different from those in the first case. According to the proposed methodology, the first bus stop of the journey (i.e., AB0110) is labelled as the boarding stop (i.e., the origin of the trip will be close to that of the bus stop). Next, for each bus stop different from AB0110, we compute the time difference between two consecutive timestamps and select the maximum value. Because it is associated with bus stop DA1268, we can infer that this last bus stop is the location where the passenger alights in the morning (i.e., the destination of the trip will be close to that bus stop) (see the left side of Fig. 10). The other bus stops represent transfer bus stops associated with the same journey.

Next, to build the OD matrix, each origin and destination bus stop is matched to each TAZ according to STEP 6. First, the distance $dist_{cf}$ between each TAZ centroid and stop/station is computed and matching it to each origin and/or destination bus stop is performed. Next, to remove the ambiguity of a single bus stop matched with two or more TAZs, the minimum distance between each TAZ centroid and the bus stop is considered. Once each bus stop is matched with each TAZ, M_3 is updated

| call code | bus stop code | route code | user code | day | time_occurrence |
|-----------|---------------|------------|--|------------|-----------------|
| 4855 | IT0549 | 8A | GsDId-002mExxFfQUaxuOky4b0eSlxUJIDkFoA4H89kgLDhNm | 21/10/2021 | 09:47:44 |
| 33847 | CF0043 | 1 | GsDId-00lgwwQNAO2QL87LPDTCGc25JNJuCtKuGnHtzP0X | 21/10/2021 | 06:52:54 |
| 21753 | HB0520 | 1 | GsDId-00Os5FdgfnF8e0IrtzfwaklXhTnIzVut110BHQINI2O | 21/10/2021 | 18:21:17 |
| 10438 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 07:56:00 |
| 10521 | AB110 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 07:57:21 |
| 10676 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 07:59:06 |
| 10688 | AB110 | 9 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 07:59:40 |
| 10992 | PM0030 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 08:19:45 |
| 11120 | PM0030 | 30 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 08:20:08 |
| 17658 | SO0020 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 16:39:00 |
| 18105 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 16:40:31 |
| 18182 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 16:44:09 |
| 19128 | DA1268 | 31 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 16:45:12 |
| 20775 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 16:56:23 |
| 21423 | RM0739 | 1 | GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheY936kihTg | 21/10/2021 | 16:57:07 |
| 30024 | DP0989 | 40 | GsDId-01Fy4Xwyk9f0ThftDGKfFDGryIXZJR0ppFsaHaRtmaJJ | 21/10/2021 | 22:08:55 |
| 30025 | DP1102 | 41 | GsDId-01Fy4Xwyk9f0ThftDGKfFDGryIXZJR0ppFsaHaRtmaJJ | 21/10/2021 | 22:07:59 |
| 41669 | DP1102 | 41 | GsDId-01Fy4Xwyk9f0ThftDGKfFDGryIXZJR0ppFsaHaRtmaJJ | 21/10/2021 | 07:54:39 |
| 46710 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7PhI9Rv0fGx2TXREwXucy | 21/10/2021 | 08:37:39 |
| 59461 | CU0741 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7PhI9Rv0fGx2TXREwXucy | 21/10/2021 | 12:59:56 |
| 59485 | DA0206 | 29 | GsDId-01Moy2EfpzLzHV2jq3X1O7PhI9Rv0fGx2TXREwXucy | 21/10/2021 | 12:59:31 |
| 48834 | P60221 | 29 | GsDId-01NgszlkNSbcPGYxzs4r6Dgc3VPNAFZP0g3xXKHHzP | 21/10/2021 | 09:04:00 |
| 39617 | LB0315 | 3 | GsDId-01W1fCakYjnahTx8wC1ItAxQM85jIleVvHkIhszVt4u | 21/10/2021 | 07:32:48 |
| 40635 | TR0311 | 3 | GsDId-01W1fCakYjnahTx8wC1ItAxQM85jIleVvHkIhszVt4u | 21/10/2021 | 07:33:03 |
| 28921 | NA0620 | 9 | GsDId-024ff6ScnKQCuzcrCikS5ER2LDYshW8nnonw6ZkWME | 21/10/2021 | 21:18:37 |
| 33099 | AF1176 | 9 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:37:55 |
| 33689 | AF1176 | 9 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 06:53:53 |
| 34611 | AF1176 | 9 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:02:26 |
| 39914 | NS0158 | 5 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:44:51 |
| 40294 | RM2109 | 5 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:45 |
| 40638 | TS0161 | 5 | GsDId-026O3BaIUxAE2Z376TDg6UZaxv5OxHrueVmTrUHD8ap8 | 21/10/2021 | 07:42:48 |
| 20976 | DD0766 | QS | GsDId-026SCyGT45IHbu0LDz652x73kqhBr0tG2L0G4GpXQA9 | 21/10/2021 | 18:12:39 |
| 21552 | BR0503 | QS | GsDId-026SCyGT45IHbu0LDz652x73kqhBr0tG2L0G4GpXQA9 | 21/10/2021 | 18:15:28 |
| 22488 | NO0701 | PQ | GsDId-026SCyGT45IHbu0LDz652x73kqhBr0tG2L0G4GpXQA9 | 21/10/2021 | 18:38:02 |
| 5094 | BD0848 | PF | GsDId-02JHndsulDGkH8Og1gUA8TK81A3Prn1qNVDI86fobgre | 21/10/2021 | 10:01:01 |

Fig. 8 A portion of M_3 : Raw data sample after the third screening algorithm (STEP 3)

with Id_{TAZ} and a single OD pair is determined for the user (see the right side of Fig. 10).

Next, by summing the related OD pairs for each user and TAZ, one can derive the corresponding matrix for each day or time period, as well as present and visualise data according to STEP 7. Figure 11 shows an example of the overall OD generated by TAZ 5510: the thicker the line, the greater the OD flow generated from the TAZ to each destination.

Interestingly, the thickest line is expected because the destination TAZ is the most common during the day, being located at a point of interest (i.e., interchange bus stops, rail, ship station, etc.). This simple application shows that the framework can be used to mine a lot of metadata on the interaction between the OD matrix and the spatial and temporal use of urban space: urban planning policies for the distribution of attractors, identification of optimal interchange parking lots, evaluation of infrastructural and preferential interventions, and organisation of the busiest nodes and stops are just a few examples. Thus, metadata would provide a synoptic picture of trends, a kind of low-cost and rapid-determination driver of the information contained in a larger and more detailed body of a universe of data.

According to STEP 8, Eqs. (6) and (7) were applied to evaluate the performance of the proposed framework. The benchmark matrix is provided by the National Institute of Statistics (ISTAT) for the area of study (ISTAT 2022). This last matrix is built by classical household travel surveys on all transport modes and refers to

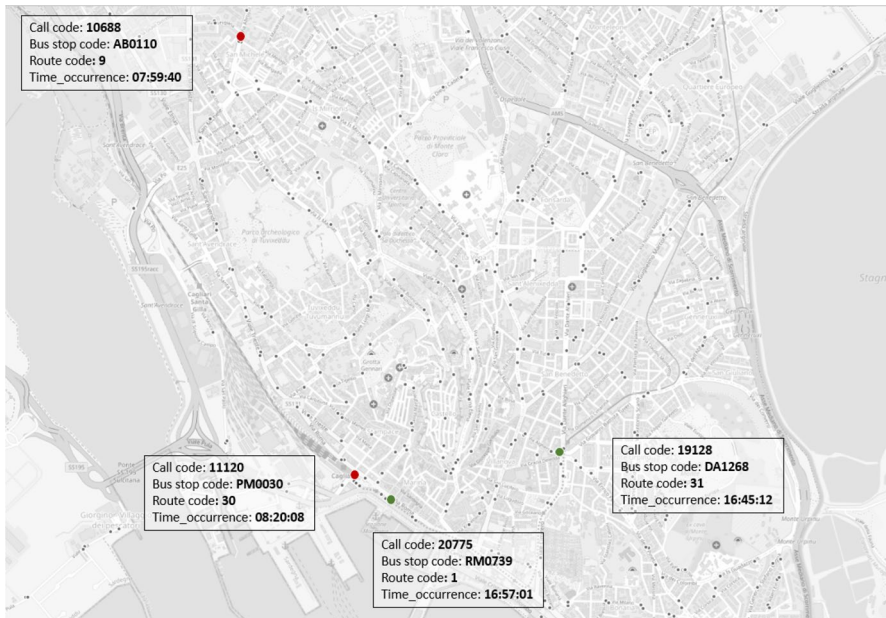


Fig. 9 Screened records of user GsDIId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg after the application of the overall data screening procedures. Red-highlighted bus stops are queried in the morning, whereas green-highlighted sections are queried during the afternoon

systematic trips only. Therefore, it was revised to refer to the bus transport mode and capture both systematic and occasional trips.

As stated before, 30 matrices with 217 TAZs are built, one for each considered workday. The results are presented in Table 4, which reports the MAE and the RMSE for both origin and destination. Note that the values of MAE and RMSE are clustered into classes for the sake of clarity. Moreover, Table 4 shows the average number of TAZs in each class and the related percentages for both origins and destinations. For instance, the second row of Table 4 shows that 116 origin TAZs and 98 destination TAZs have a high accuracy in estimations, because the MAE is lower than 1.

The overall results show that:

- MAE is not very high. For instance, for about 90% of the origin trips and 85% of the destination trips, the trip error rate is lower than 5 movements and looks very accurate. The average MAE is about 2.31 trips, which is very low.
- $RMSE \neq MAE$, thus all errors are not of the same magnitude.
- $RMSE > MAE$, thus there is some variation in the magnitude of the errors.
- The difference between RMSE and MAE is not so large, thus very large errors are unlikely to have occurred.

Therefore, the proposed method that uses SACDRs data seems accurate for inferring OD matrices.



Fig. 10 (Top): OD of user GsDId-04kcsvU5T3JQaAsaxUuVx7uGpJG3etdAheYY936kihTg; (Down): the trip of the same user after the application of algorithm in STEP 6; note that black numbers represent c^* , i.e., the label of TAZs's centroid

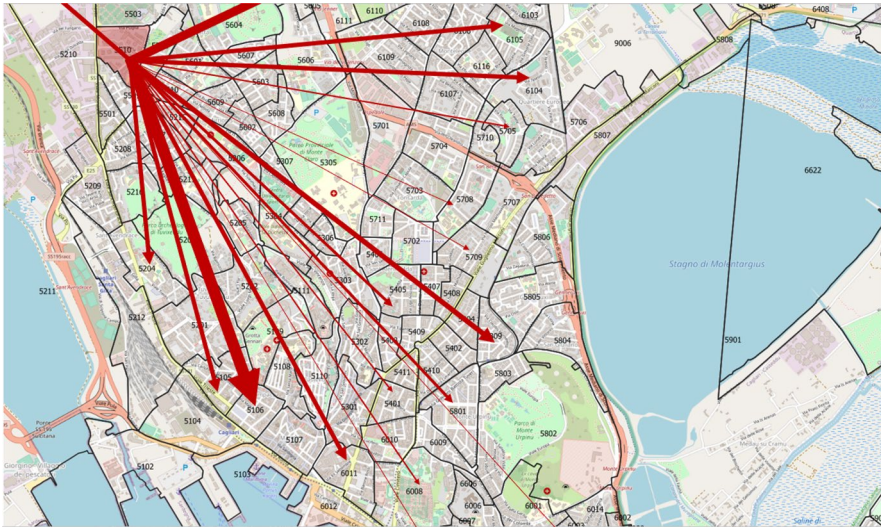


Fig. 11 OD flow generated from TAZ 5510 to each destination. The black numbers represent c^* , i.e., the label of TAZs's centroid

Table 4 Average errors over the 30 workdays

| | Average number of origins (TAZ) | Average percentage | Average number of destinations (TAZ) | Average percentage |
|---------------------|---------------------------------|--------------------|--------------------------------------|--------------------|
| <i>MAE [trips]</i> | | | | |
| 0.00–1.00 | 114 | 52.33% | 97 | 44.73% |
| >1.00–5.00 | 80 | 37.05% | 88 | 40.34% |
| >5.00–10.00 | 9 | 4.21% | 25 | 11.31% |
| >10.00 | 14 | 6.41% | 8 | 3.63% |
| <i>RMSE [trips]</i> | | | | |
| 0.00–2.00 | 78 | 35.87% | 65 | 29.82% |
| >2.00–10.00 | 103 | 47.65% | 93 | 43.07% |
| >10.00–20.00 | 13 | 6.04% | 30 | 13.93% |
| >20.00 | 23 | 10.45% | 29 | 13.18% |

As a further check, the $r_{ISTAT,SACDR}$ is computed by Eq. (8) to evaluate the correlation between the ISTAT matrix and the SACDR matrix. Figures 12 and 13 show the $r_{ISTAT,SACDR}$ for each TAZ between SACDRs and ISTAT data, for origin and destination trips, respectively.

Both figures show a high correlation mainly in the first and second municipalities of Cagliari (red dots and, to a lesser extent green dots), whereas the correlation is acceptable in the central business district, which has few residents due to high costs and the presence of many service buildings, administration, schools, etc. On the average, the $r_{ISTAT,SACDR}$ at the origin takes a value of 70%, whereas at the

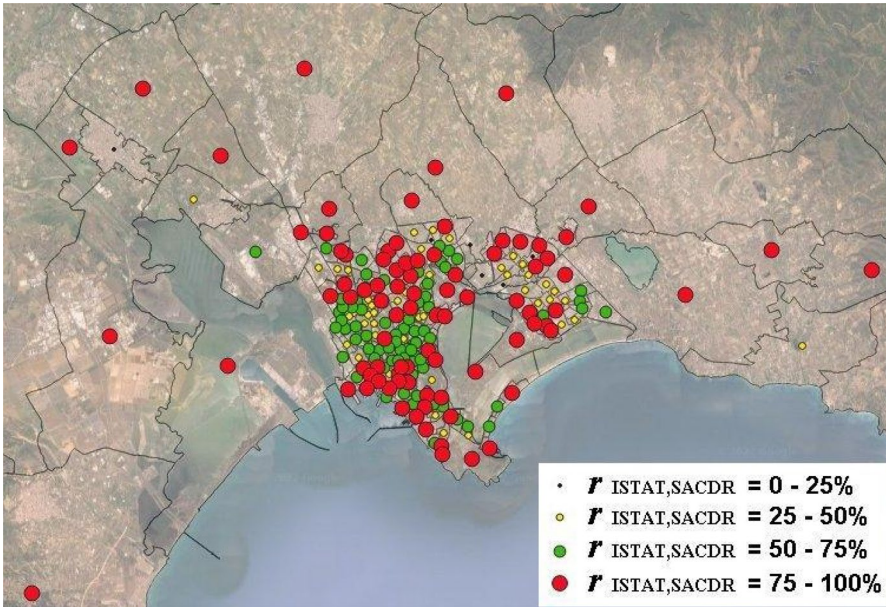


Fig. 12 $r_{\text{ISTAT,SACDR}}$ for origin trips between ISTAT and SACDR matrices

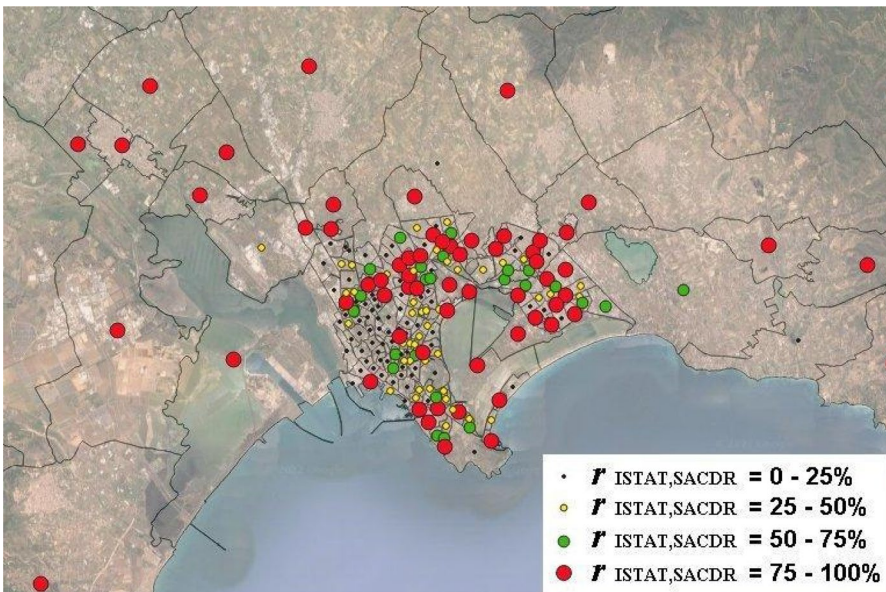


Fig. 13 $r_{\text{ISTAT,SACDR}}$ for destination trips between ISTAT and SACDR matrices

destination it is larger than 50%. Thus, the correlation is high, and the proposed method provides a quite reliable estimation of OD trips using buses.

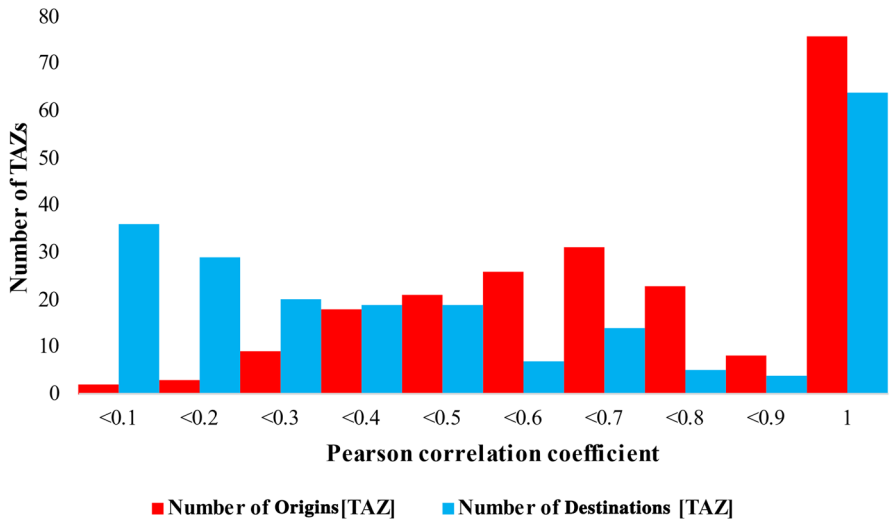


Fig. 14 Distribution of $r_{ISTAT,SACDR}$ for origin and destination TAZs

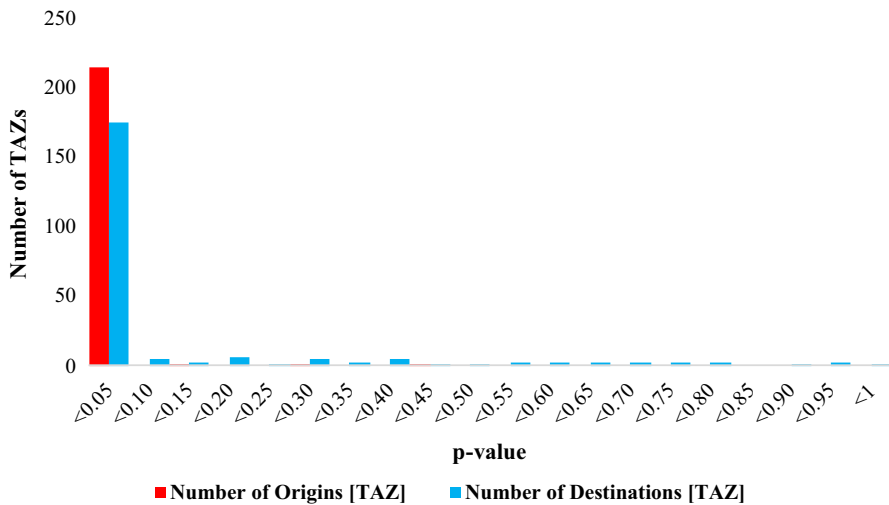


Fig. 15 Distribution of p -value for origin and destination TAZs

Moreover, Fig. 14 shows the distributions of $r_{ISTAT,SACDR}$ for origin and destination TAZs. It shows that the correlation is higher for origin TAZs. Indeed, about 85% of the origin TAZs have a coefficient of correlation higher than 0.5, as opposed to about 53% of the destination TAZs.

Finally, 217 LRMs for each specific origin TAZ, and 217 LRMs for each specific destination TAZ, are fitted according to Eq. (9). Figure 15 shows the results for the

statistical significance of each LRM, according to the F-test. It indicates that a substantial portion of the LRMs is very significant, although the significance is higher for the origin TAZs. Indeed, about 99 and 80% of origin and destination TAZs have a *p-value* lower than 0.05, respectively.

Both figures indicate a strong relationship between the target and the predicted number of trips, because all the Linear Regression Models are very significant according to the F-Test.

To sum up, even if SACDRs data return a slightly different matrix from ISTAT, the interest of the SACDRs method is evident, because an OD matrix could be obtained daily with lower effort as opposed to traditional approaches. Conversely, the ISTAT matrix is an 'average' matrix that needs frequent updating. Thus, the validation enabled comparing the OD SACDR matrix with the ISTAT data with promising results.

5 Conclusions and research perspectives

The knowledge of passenger demand is a major requirement for public transport companies (PTCs). Although it can be estimated by traditional methods based on surveys, ride-check counts and smartcards, the availability of new communication technologies and the rapid spread of portable devices has paved the way for emerging methods to support this activity by telephone networks, Wi-Fi signals and apps. However, the challenge of processing collected data must be faced to derive meaningful information.

This paper proposes the first framework for the estimation of OD matrices by processing the data provided by the app of a bus operator. The paper shows how to (i) perform some screening activities to process relevant SACDRs data, (ii) link the SACDRs data of all passengers to reconstruct their journey, (iii) infer the origin and destination of each passenger for the period of interest and (iv) validate the estimated flows. More precisely, the framework:

- Integrates four novel algorithms for handling large amounts of SACDRs data to infer OD matrices in public transport over the overall transit network.
- Generates a mainstream source of origin and destination flows automatically, to shed light on which ODs need proper care because of the magnitude of their flows.
- Is illustrated in a real case study to show its practical effectiveness. Appropriate maps were constructed to plan possible improvements in the service of bus routes once OD matrices are inferred. Moreover, the results show a quite reliable estimation w.r.t. a benchmarking matrix.

This study expands upon the ideas presented in Obino et al. (2023), incorporating additional data and analyses for a more comprehensive statistical examination.

There are several remarkable implications in the use of this framework for demand estimation. Unlike smart cards, the outcomes of this method are not affected by fare evasion and the availability of multiple types of tickets. Unlike long-term

based surveys, this method enables to build daily OD matrices, thus providing PTCs with a much deeper knowledge of passenger demand. Moreover, the framework returns OD matrices at much lower costs than survey-based, smartcard-based, and portable device-based methods, because it avoids costly surveys and/or sensors and/or other devices on board or in ground infrastructures. Yet, since the framework can analyse a source of endogenous data on passenger demand, it results in significant workload savings for the planning department of PTCs. Finally, data privacy is preserved by an anonymized use-code.

The method can be extended to include incremental features to limit possible shortcomings. First, although the experimentation was done by a PTC in a monopolistic setting, other competitors must adopt a similar app and share collected data to avoid demand underestimation. Second, the scale of this research could be extended beyond the case of a mid-sized Italian PTC. For example, the framework is tested in the case only by buses and trolleys, whereas other transit modes could be considered (e.g., underground, light rail transit occurring in big cities). Finally, comparative experiments among this framework and the smartcard-based methods and/or portable device-based methods cannot be currently tested in this specific study owing to the lack of benchmarks and budget constraints.

Notwithstanding, additional developments will be investigated. First, the framework may be integrated with other data sources, such as automatic passenger counting and automatic fare collection systems to improve the estimation of OD passengers. Thus, new data fusion and data handling algorithms should be implemented to make this integration. Second, even if running times were not a key point of this paper, it is important to investigate the computation efficiency of the framework in the case of large amounts of data. Third, the framework can be applied every day to derive several OD bus matrices, arrange service frequency and revise bus routes accordingly, to improve service quality. The estimation of passenger demand across different time periods with a focus on peak hours is relevant for enhancing public transport services. It can enable better resource allocation, improve service planning, enhance commuter experiences, reduce crowding on board and ensure smoother and more reliable journeys for passengers. The potential integration of different data sources and the incorporation of advanced modelling techniques will help create responsive and sustainable transit systems adapting to the dynamic needs of urban mobility.

Future research will investigate estimations at the route level, which may lead to the improvement of the service design. For instance, the route-level passenger OD flow can provide valuable information for the determination of new stop locations, route changes (e.g., extension, splitting, or merging) and the introduction of new services. Moreover, the route-level OD matrix can be useful for investigating the crowding on board along the route to improve service quality. These research topics may greatly influence future smart cities (Garau et al. 2022).

Acknowledgements The authors are grateful to the General Management of CTM SpA for its support in this research and the possibility of using its real data for experimentation. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the CTM. Any remaining errors are the authors' responsibility alone. This study was supported by the Ministry of Education, Universities and Research (MIUR) (Italy) through

a project entitled WEAKI TRANSIT: WEAK-demand areas Innovative TRANSPORT Shared services for Italian Towns (Project code: 20174ARRHT /CUP Code: E44117000050001), financed by the PRIN 2017 (Research Projects of National Relevance) program. We authorise the MIUR to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright notations thereon. Moreover, this research was partially funded by Fondazione di Sardegna, Progetto biennale bando 2021, “Computational Methods and Networks in Civil Engineering (COMANCHE)”. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the MIUR. Finally, this work was partially funded by Department of Civil, Environment, Land and Architecture Engineering and Mathematics (DICATAM), University of Brescia through the research grant “Valuation of the risk of fare evasion in an urban public transport network”, CUP: D73C22000770002.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement.

Data availability The authors do not have permission to share data.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- American Public Transportation Association—APTA (2021) Transit ridership report fourth quarter and end-of-year 2017 (PDF). Archived from the original (PDF) on 27 Mar 2018. Retrieved 10 Oct 2021
- Barabino B, Di Francesco M, Ventura R (2023) Evaluating fare evasion risk in bus transit networks. *Transp Res Interdiscip Perspect* 20:100854
- Barabino B, Lai C, Olivo A (2020) Fare evasion in public transport systems: a review of the literature. *Public Transp* 12(1):27–88. <https://doi.org/10.1007/s12469-019-00225-w>
- Barabino B, Di Francesco M, Mozzoni S (2014) an offline framework for handling automatic passenger counting raw data. *IEEE Trans Intell Transp Syst* 15(6):2443–2456
- Ben-Akiva M, Macke P, Hsu P (1985) Alternative methods to estimate route-level trip tables and expand on-board surveys. *Transp Res Rec* 1037:1–11
- Blum J, Sridhar A, Mathew T (2010) Origin–destination matrix generation from boarding–alighting and household survey data. *Transp Res Rec* 2183:1–8. <https://doi.org/10.3141/2183-01>
- Caceres N, Wideberg J, Benitez F (2008) Deriving origin destination data from a mobile phone network. *Intell Transp Syst IET* 1(1):15–26
- Calabrese F, Lorenzo G, Liu L, Ratti C (2011) Estimating origin–destination flows using mobile phone location data. *IEEE Pervasive Comput* 10(4):36–44
- Carrel A, Lau PS, Mishalani RG, Sengupta R, Walker JL (2015) Quantifying transit travel experiences from the users’ perspective with high-resolution smartphone and vehicle location data: methodologies, validation, and example analyses. *Transp Res Part C Emerg Technol* 58:224–239
- Chapin N (1970) Flowcharting with the ANSI standard: a tutorial. *ACM Comput Surv (CSUR)* 2(2):119–146
- Chaudhary M, Bansal A, Bansal D, Raman B, Ramakrishnan KK, Aggarwal N (2016) Finding occupancy in buses using crowdsourced data from smartphones. In: Proceedings of the 17th international conference on distributed computing and networking. ACM, pp 35–39

- Chu K, Chapeau R (2010) Augmenting transit trip characterization and travel behavior comprehension: multiday location-stamped smart card transactions. *Transp Res Rec* 2183:29–40. <https://doi.org/10.3141/2183-04>
- CTM (2020) Carta della mobilità 2019–2020. <https://www.ctmcagliari.it/>
- De Aloe M, Ventura R, Bonera M, Barabino B, Maternini G (2023) Applying cost-benefit analysis to the economic evaluation of a tram-train system: evidence from Brescia (Italy). *Res Transp Bus Manag* 47:100916
- Demissie MG, Phithakkitnukoon S, Sukhvibul T, Antunes F, Gomes R, Bento C (2016) Inferring passenger travel demand to improve urban mobility in developing countries using cell phone data: a case study of Senegal. *IEEE Trans Intell Transp Syst* 17(9):2466–2478
- Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. *Proc Nat Acad Sci* 111(45):15888–15893
- Drosouli I, Voulodimos A, Miaoulis G, Mastorocostas P, Ghazanfarpour D (2021) Transportation mode detection using an optimized long short-term memory model on multimodal sensor data. *Entropy* 23:1457
- Furth PG, Navick DS (1992) Bus route OD matrix generation: relationship between biproportional and recursive methods. *Transp Res Rec* 1338:14–21
- Garau C, Desogus G, Barabino B, Coni M (2022) Accessibility and public transport mobility for a smart(er) island: evidence from Sardinia (Italy). *Sustain Cities Soc* 87:104145
- Ge L, Sarhani M, Voß S, Xie L (2021) Review of transit data sources: potentials, challenges and complementarity. *Sustainability* 13(20):11450. <https://doi.org/10.3390/su132011450>
- Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779–782. <https://doi.org/10.1038/nature06958>
- Gonzalez ABR, Diaz JJV, Wilby MR (2020) Detailed origin–destination matrices of bus passengers using radio frequency identification. *IEEE Intell Transp Syst Mag* 14(1):141–152
- Håkegård JE, Myrvoll TA, Skoglund TR (2018) Statistical modelling for estimation of OD matrices for public transport using Wi-Fi and APC data. In: 21st International conference on intelligent transportation systems (ITSC), pp 1005–1010
- Hazelton ML (2010) Statistical inference for transit system origin-destination matrices. *Technometrics* 52(2):221–230
- Iqbal MS, Choudhury CF, Wang P, Gonzalez MC (2014) Development of origin destination matrices using mobile phone call data. *Transp Res Part C Emerg Technol* 40:6374
- ISTAT (2022) <https://www.istat.it/>. Accessed 31 Aug 2022
- Ji Y, Mishalani RG, McCord MR (2015a) Transit passenger origin-destination flow estimation: efficiently combining onboard survey and large automatic passenger count datasets. *Transp Res Part C Emerg Technol* 58:178–192
- Ji Y, You Q, Jiang S, Zhang HM (2015b) Statistical inference on transit route-level origin-destination flows using automatic passenger counter data. *J Adv Transp* 49(6):724–737
- Jee H, Sun W, Schmöcker JD, Nakamura T (2023) Demonstrating the feasibility of using Wi-Fi sensors for dynamic bus-stop queue length estimation. *Public Transp* 1–18. <https://doi.org/10.1007/s12469-023-00336-5>
- 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. Master’s in science Geomatics
- Li Y, Cassidy MJ (2007) A generalized and efficient algorithm for estimating transit route ODs from passenger counts. *Transp Res Part B Methodol* 41(1):114–125
- Liu D, Zhao C, Dong H, Huang Z (2022) Spatial analysis of bus rapid transit actual operating conditions: the case of Hangzhou City, China. *Public Transp* 14:503–519. <https://doi.org/10.1007/s12469-022-00299-z>
- Liu H, Yang J, Sidhom S, Wang Y, Chen Y, Ye F (2013) Accurate WiFi based localization for smartphones using peer assistance. *IEEE Trans Mob Comput* 13:2199–2214
- Lu Y, Misra A, Sun W, Wu H (2017) Smartphone sensing meets transport data: a collaborative framework for transportation service analytics. *IEEE Trans Mob Comput* 17:945–960. <https://doi.org/10.1109/TMC.2017.2743176>
- McCord MR, Mishalani RG, Goel P, Strohl B (2010) Iterative proportional fitting procedure to determine bus route passenger origin-destination flows. *Transp Res Rec* 2145(1):59–65

- Mishalani RG, McCord MR, 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(1):123–130
- Mishalani RG, Ji Y, McCord MR (2011) Effect of onboard survey sample size on estimation of transit bus route passenger origin-destination flow matrix using automatic passenger counter data. *Transp Res Rec* 2246(1):64–73
- Munizaga MA, Palma C (2012) Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. *Transp Res Part C Emerg Technol* 24:9–18
- Munizaga MA, Gschwender A, Gallegos N (2020) Fare evasion correction for smartcard-based origin-destination matrices. *Transp Res Part A Policy Pract* 141:307–322
- Murray AT, Davis R, Stimson RJ, Ferrera J (1998) Public transportation access. *Transp Res Part D Transp Environ* 3(5):319–328
- Myrvoll TA, Håkegård JE, Matsui T, Septier F (2017) Counting public transport passenger using WiFi signatures of mobile devices. In: *IEEE 20th international conference on intelligent transportation systems (ITSC)*, pp 1–6
- Navick DS, Furth PG (1994) Distance-based model for estimating a bus route origin-destination matrix. *Transp Res Rec* 1433:16–23
- Nitti M, Pinna F, Pintor L, Pilloni V, Barabino B (2020) iABACUS: a wi-fi-based automatic bus passenger counting system. *Energies* 13(6):1446
- Oberli C, Torres-Torriti M, Landau D (2010) Performance evaluation of UHF RFID technologies for real-time passenger recognition in intelligent public transportation systems. *IEEE Trans Intell Transp Syst* 11(3):748–753
- Obino A, Ventura R, Coni M, Di Francesco M, Barabino B (2023) Building origin-destination matrices in bus networks from smartphone-app call detail records: evidence from Italy. In: Presented at XXVI international conference LWC 2023, Brescia (Italy), 6–8 September 2023
- Olivo A, Maternini G, Barabino B (2019) Empirical study on the accuracy and precision of automatic passenger counting in European bus services. *Open Transp J* 13(1):250–260
- Ortuzar JDD, Willumsen LG (2011) *Modelling transport*, 3rd edn. West Sussex, Wiley
- Pelletier MP, Trépanier M, Morency C (2011) Smart card data use in public transit. A literature review. *Transp Res Part C Emerg Technol* 19:557–568
- Phithakkitnukoon S, Veloso M, Bento C, Biderman A, Ratti C (2010) Taxi-aware map: identifying and predicting vacant taxis in the city. *Int Jt Conf Ambient Intell* 6439:86–95
- Rahbee AB (2008) Farecard passenger flow model at Chicago transit authority, Illinois. *Transp Res Rec* 2072(1):3–9
- Rossetti MD, Turitto T (2000) Design of an integrated transit monitoring system based on radio frequency identification. *Int J Serv Technol Manage* 1(2–3):188–204
- Seaborn C, Attanucci J, Wilson N (2009) Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transp Res Rec* 2121:55–62. <https://doi.org/10.3141/2121-06>
- Simon J, Furth PG (1985) Generating a bus route OD matrix from on-off data. *J Transp Eng* 111(6):583–593
- Tamin OZ (1997) Public transport demand estimation by calibrating a trip distribution mode choice (TDMC) model from passenger counts: A case study in Bandung, Indonesia. *J Adv Transp* 31(1):5–18
- Tilocca P, Farris S, Angius S, Argiolas R, Obino A, Secchi S, Mozzoni S, Barabino B (2017) Managing data and rethinking applications in an innovative mid-sized bus fleet. *Transp Res Procedia* 25:1899–1919
- Trépanier M, Tranchant N, Champleau R (2007) Individual trip destination estimation in a transit smart card automated fare collection system. *J Intell Transp Syst* 11(1):1–14
- Tu L, Wang S, Zhang D, Zhang F, He T (2019) ViFi-MobiScanner: observe human mobility via vehicular internet service. *IEEE Trans Intell Transp Syst* 22(1):280–292
- Ventura R, Bonera M, Carra M, Barabino B, Maternini G (2022) Evaluating the viability of a tram-train system. A case study from Salento (Italy). *Case Stud Transp Pol* 10(3):1945–1963. <https://doi.org/10.1016/j.cstp.2022.08.009>
- Wang W, Attanucci J, Wilson NH (2011) Bus passenger origin–destination estimation and related analyses using automated data collection systems. *J Public Transp* 14(4):131–150
- Zhang L, Zhao S, Zhu Y, Zhu Z (2007) Study on the method of constructing bus stops OD matrix based on IC card data. In: *IEEE international conference on wireless communications, networking and mobile computing*, pp 3147–3150

- Zhao J, Rahbee A, Wilson NH (2007) Estimating a rail passenger trip origin–destination matrix using automatic data collection systems. *Comput Aided Civ Infrastruct Eng* 22(5):376–387
- Zúñiga F, Muñoz JC, Giesen R (2021) Estimation and prediction of dynamic matrix travel on a public transport corridor using historical data and real-time information. *Public Transp* 13:59–80. <https://doi.org/10.1007/s12469-020-00255-9>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Benedetto Barabino¹ · Mauro Coni² · Massimo Di Francesco³  · Andrea Obino⁴ · Roberto Ventura¹

✉ Massimo Di Francesco
mdifrance@unica.it

- ¹ Department of Civil, Environment, Land and Architecture Engineering and Mathematics, University of Brescia, Brescia, Italy
- ² Department of Civil, Environment and Architecture Engineering (DICAAR), University of Cagliari, Cagliari, Italy
- ³ Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy
- ⁴ Department of General Management, CTM SpA, Cagliari, Italy