A data-driven approach for origin-destination matrix construction from cellular network signalling data: a case study of Lyon region (France)

Mariem Fekih, Tom Bellemans, Zbigniew Smoreda, Patrick Bonnel, Angelo Furno, Stéphane Galland

A data-driven approach for Origin-Destination matrix construction from cellular network signalling data: a case study of Lyon region (France)

Mariem Fekih [a,b,1], Tom Bellemans [a], Zbigniew Smoreda [b], Patrick Bonnel [c], Angelo Furno [d], Stéphane Galland [e]

[a] *Transportation Research Institute (IMOB), Hasselt University, Wetenschapspark 5 bus 6, 3590 Diepenbeek, Belgium*
[b] *SENSE, Orange Labs, 92326 Chatillon Cedex, France*
[c] *LAET, ENTPE, 69518 Vaulx-en-Velin Cedex, France*
[d] *LICIT, ENTPE, Univ. Lyon, Univ. Gustave Eiffel, F-69675 Bron Cedex, France*
[e] *CIAD, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France*

Abstract

Spatiotemporal data, and more specifically origin-destination matrices, are critical inputs to mobility studies for transportation planning and urban management purposes. Traditionally, high-cost and hard-to-update household travel surveys are used to produce large-scale origin-destination flow information of individuals' whereabouts.
In this paper, we propose a methodology to estimate Origin-Destination (O-D) matrices based on passively-collected cellular network signalling data of millions of anonymous mobile phone users in the Rhône-Alpes region, France. Unlike Call Detail Record (CDR) data which rely only on phone usage, signalling data include all network-based records providing higher spatiotemporal granularity. The explored dataset, which consists of time-stamped traces from 2G and 3G cellular networks with users' unique identifier and cell tower locations, is used to first analyse the cell phone activity degree indicators of each user in order to qualify the mobility information involved in these records. These indicators serve as filtering criteria to identify users whose device transactions are sufficiently distributed over the analysed period to allow studying their mobility. Trips are then extracted from the spatiotemporal traces of users for whom the home location could be detected. Trips have been derived based on a minimum stationary time assumption that enables to determine activity (stop) zones for each user. As a large, but still partial, fraction of the population is observed, scaling is required to obtain an O-D matrix for the full population. We propose a method to perform this scaling and we show that signalling data-based O-D matrix carries similar estimations as those that can be obtained via travel surveys.

Keywords: Passive cellular signalling data, Big data analysis, travel survey, home detection, trip extraction, origin-destination matrix

1. Introduction

Spatiotemporal data are extremely valuable to study human mobility for transportation and urban planning purposes (Arentze and Timmermans 2000; Giannotti and Pedreschi 2008). At large scale (e.g. regional level), traditional approaches rely on household travel surveys to collect mobility data that typically record one day of travel diaries per household. While travel surveys provide highly useful data to formalize and estimate behavioural transport

[1] Corresponding author. Tel.: +33 6 03 41 05 66
E-mail address: mariem.fekih@uhasselt.be (Mariem Fekih)

models (e.g. route and transportation mode choice models), they are much less useful for constructing origin-destination (O-D) matrices due to limited sample sizes, which result in empty cells in the matrix estimation. Indeed, surveys are increasingly confronted by issues during the sample construction phase (Stopher and Greaves 2007), by declining response rates (Bonnel 2003) and by unreported trips (Wolf et al. 2003), which reduce even further the quality of the resulting matrices. Additionally, travel surveys typically involve high costs that restrict their frequency (once or twice per decade) and prevent to follow the dynamics of population mobility over time.

Several kinds of sensor data dealing with the position and mobility of individuals have become recently available due to the wide deployment of pervasive computing equipment. Hence, large volumes of data are being produced automatically and passively from different technologies, such as GPS based-devices, smart cards and mobile phones, which make it possible to identify the presence of individuals in both space and time (Feng and Timmermans 2014; Munizaga and Palma 2012). In particular, data collected from cell phones have become one of the most important new data sources to study travel behaviour (Wang et al. 2017). Their proper attributes, such as large coverage of geographic area, significant penetration in population and high detailed location information have attracted researchers to analyse them to support transportation studies. A number of researches have been conducted to use different types of mobile phone data (e.g., Call Detail Records (CDR), cellular network data); but, few have attempted to validate the results with external sources due to the different nature of mobile phone footprints. Yet, the validation process allows to identify possible biases and to gain a clearer idea of their potential. Moreover, the quality and accuracy of data is essential to ensure that investment or transport policy decisions are based on reliable analyses. Therefore, considerable efforts are needed to pre-process mobile phone data and to validate the related research outputs.

The aim of this paper is to explore cellular signalling data from 2G and 3G networks to produce origin-destination matrices. Although the potential of these data is promising due to the involved large amounts of individual spatiotemporal traces comparing to CDR data, there is still a remarkable lack of studies based on them.

Our primary goal is to test whether these massive signalling data could act as cheap and reliable data source to capture individual trips. Therefore, we propose, as the main contribution of this paper, a full workflow to transform cell phone network logs into origin-destination flow matrices supported by a validation step using travel survey data. Our approach is evaluated in a case study related to the Rhône-Alpes region, France, for which we were able to analyse recent mobile phone signalling data (June 2017) provided by Orange, the largest French mobile operator, and compare them with the data obtained from the latest travel survey performed by the local authority in the same region.

This paper is structured as follows. Section 2 describes the related work. In Section 3, an overview of the data used in our analysis is presented. In Section 4, the methodology applied to estimate the O-D matrix from signalling data is discussed. While in Section 5, our results are summarized and validated with respect to travel survey data. Section 6 discusses our study contribution to the traditional survey methods and with respect to the existing works as well as our signalling data-based approach challenges. Finally, Section 7 concludes the paper and identifies several suggestions for future research directions.

2. Related works

2.1. Mobile phone data for travel behaviour research

The wide adoption of mobile devices (mobile phones, smartphones and tablets) and the rapid related advancements make the mobile phone data a good candidate for the study of human mobility for transportation research. Indeed cell phone networks have existed for three decades, and mobile phones have achieved a high rate of penetration: there were 75 million active SIM (Subscriber Identity Module) cards in France in 2018, for a total population of 66 million (ARCEP 2018). The exploration of the large and passive datasets generated from cellular networks seems to have enormous potential in the field of travel behaviour studies.

Hereby, mobility modelling has experienced rapid developments in recent years thanks to these new sensor data. González et al. (2008) have proposed one of the first studies of large-scale mobility using a Call Detail Records (CDR) sample of over 100,000 users to explore the universal laws of individual human mobility. This study demonstrated that the distribution of users' trips is well approximated by a truncated power-law distribution. More recently, mobile

phone data have been explored for mobility pattern extraction (Asgari et al. 2013; Calabrese et al. 2013; Wang et al. 2012), traffic and mobility flows inference (Calabrese et al. 2011a; Huang et al. 2018) , population estimation (Frias-Martinez et al. 2010; Ricciato et al. 2016) and route choice modelling (Tettamanti and Istvan 2014).

Moreover, terminal logs have been explored to detect individual activities. Widhalm et al. (2015) have developed activity program typologies based on duration analysis, trips and activity location and frequencies combined with spatial typologies. They applied the method in the cities of Vienna and Boston showing similarities between conurbations but also some local specificities. Xu et al. (2015) have studied spatial distribution of individual activity area from home in Shenzen, China. Jiang et al. (2016) applied the stay extraction method defined in (Jiang et al. 2013) to infer stay location type. They were able to identify home locations for 75% of users based on the record start time and the frequency of visits to each stay location. Accordingly, Jiang et al. (2017) used CDR data to model activity patterns of identified resident subscribers in Singapore. In a study using simulated sample from device-based location data and household travel survey, Chen et al. (2014) suggested a set of methods to detect activity locations and their types. In this way, they have shown that the presented procedure reproduces individuals' home and work with fairly high degree of accuracy, and, with less accuracy the location of the places they visit.

Data from the mobile phone network can also be used to estimate individual trajectories. Schlaich et al. (2010) developed an algorithm that was able to identify precisely a user's trajectory between the cities of Karlsruhe and Stuttgart in Germany, considering mainly the "location-area-sequences" events. Similarly, road traffic monitoring, i.e. route detection, has been investigated using signalling data by (Fiadino et al. 2012). In 2014, other research effort has focused on human trajectory extraction using the interpolation methods (Hoteit et al. 2014).

As an additional challenge, mobile phone signalling data has been applied to infer travel modes by estimating coarse speeds according to the change rates of connected cells and the fluctuations of signal strength (Feng and Timmermans 2014).The estimation accuracy is quite high. However, the identifiable travel modes are still rather limited with this kind of data. Furthermore, cell network traces were used to help deriving critical transport-related measures such as mean speeds, travel distance and journey times. Among recent studies, Calabrese et al. (2013; 2011b), working in the Boston conurbation, used triangulated mobile phone data collected by a telecom operator to study mean speed, mean trip length and the distribution according to the time of day. The very recent research conducted by Janzen et al. (2018) concerns particularly the analysis of the long-distance trips carried out over the entire area of France. It illustrates the considerable potential of mobile phone CDR data for the analysis of long-distance trips.

GPS (Global Positionning System) data represent an alternative form of mobile phone data largely leveraged in active travel research. Owing to their ubiquity, GPS-enabled mobile phones can be used to collect GPS records (Gonzalez et al. 2010; Nitsche et al. 2014; Nour et al. 2016; Widhalm et al. 2012) and largely decrease the data collection cost comparing to the dedicated GPS loggers used in GPS-based surveys (Bohte and Maat 2009; Deutsch et al. 2012; Stopher et al. 2008; Wolf et al. 2004). As for typical GPS data, different methods for different purposes could be adopted (Shen and Stopher 2014). Rule-based algorithms have been mostly used to extract stays and trips by setting an activity duration threshold (e.g. 120 seconds) between consecutive records (Choujaa 2009; Feng and Timmermans 2014). Rule-based approaches have also been used to impute activities and trip purposes from GPS traces using GIS (Geographic Information System) land use data (Bohte and Maat 2009; Chen et al. 2010; Stopher et al. 2008) by considering several measures such as information about surrounding point of interests (Huang et al. 2010). Classification techniques and learning-based systems have been primarily applied for travel mode detection: the main feature used for these approaches is travel speed supplemented with transport network information. (Reddy et al. 2010) used built-in GPS and accelerometer data from cell phones to train a classifier system consisting in a decision tree followed by a discrete Hidden Markov Model. In the experiment, 16 volunteers were asked to carry six phones positioned on different places for 15 minutes for each mode. The proposed approach is able to identify transportation modes, including stationary, walking, running, biking, or motorized modes with an accuracy of 93.6%. In (Gonzalez et al. 2010), the authors use GPS data recorded via a custom mobile phone application to investigate the feasibility of automatic travel mode detection with neural networks (NNs). The focus is to provide a convenient technique that uses a minimum set of GPS fixes in order to save device resources (e.g. battery life, data transfer costs etc.) during the data collection process. Although GPS data successfully allow to reduce the number of underreported trips and to improve the estimation accuracy of traditional surveys, they still have major limitations. First, this kind of data is either collected from specific smartphone applications (e.g. using built-in GPS receiver) developed for research purposes or

from dedicated GPS devices (e.g. person-based GPS logger) resulting in additional burden to the participant. This widely restricts the sample size since it is not easy to recruit a sufficient number of participants (from tens to few hundred surveyed individuals in existing studies). Second, these data suffer from significant sampling bias due to the selected random group of persons (Nitsche et al. 2014). Moreover, GPS data collection suffers from technical issues such as signal noise and signal loss as well as the well-known data storage problem (in case of on-device storage) and data transfer cost which are still challenging the quality and relevance of GPS fixes. Thus, unlike cellular network data, these survey methods appear to be not well suited for longitudinal travel surveys and large-scale mobility behavior studies.

An illustrative overview of some state-of-the-art works in the field of urban and travel behaviour analysis using mobile phone traffic data is summarized in (Calabrese et al. 2014; Blondel et al. 2015; Naboulsi et al. 2016; Wang et al. 2017).

2.2. Origin-Destination estimation using cellular network-based data

Transport planners mostly rely on transport demand models for the understanding of mobility behaviour and the planning of network infrastructure (Bonnel 2004; Ortúzar and Willumsen 2011). These transport travel demand models are heavily relying on high-cost and hard-to-update travel surveys as a data source. Indeed, these traditional collection methods result in an overview of the mobility of one weekday; thereby they only provide a snapshot of people movement since they cover a limited sample of population and small time window (Nitsche et al. 2014). Therefore, demand models could not be updated regularly and so might not reflect real mobility behaviour which results to the need of new data sources.

The last years have witnessed a surge of studies using passively generated traces from cellular devices to estimate origin-destination flows. Furthermore, there have been several limited-scale researches aimed at analysing the potential of these emerging data for origin-destination matrix construction. In 2002, a small sample from one morning has been used to study traffic O-D matrices on specific roads in the county of Kent in the UK (White and Wells 2002). Authors revealed the potential of billing data and suggested that more research was needed to infer consistent O-D matrix. Later in 2007, Caceres et al. (2007) calculated an O-D matrix with four possible O-D pairs to study the road traffic in the highway between the cities of Huelva and Seville in Spain. They employed simulated mobile phone data and compared the results with those obtained from traffic counts showing again the potential of such cost-effective method. Both of these studies are based on too small samples of CDR covering very limited areas.

More recently, Calabrese et al. (2011b) were the first to produce an O-D matrix from a detailed mobile phone dataset (e.g. including more internet connections), for the Boston region in Massachusetts, showing encouraging comparison results with O-D flows from census data taking account of only weekday morning trips. Mellegard et al. (2011) adapted an algorithm method to the available database of mobile device records that covers a large territory of Sweden to generate O-D flows. No detailed comparison for the entire matrix has been performed in this work. Moreover, to derive travel demand and routes, CDR data have been explored to generate “transient O-D matrices” and to convert them into intersection-to-intersection O-D flows in the road network of Boston and San Francisco (Wang et al. 2012) and in Dhaka, Bangladesh (Iqbal et al. 2014). To calibrate the derived O-D trips, (Wang et al. 2012) have used available travel data, vehicle usage rate and population statistics. High correlation was identified when validating the up-scaled flows with probe vehicle GPS data. Using limited traffic counts and a microscopic simulation, (Iqbal et al. 2014) have scaled up the generated OD patterns (from calling data). They have validated the assignment results with additional traffic counts (different to those for calibration) and found that the prediction error was about 13%. The limited availability of high resolution GIS travel data and sufficient amount of traffic counts (e.g. due to their collection cost) makes these methodologies less applicable especially in case of large-scale studies (e.g. regional or national). Alexander et al. (2015) have conducted analysis on triangulated CDR data (with estimated *(x, y)* device coordinates) to infer O-D individual trips per purpose (home, work or other) and time of the day. After a filtering process, they kept only about 16% of users to extract trips. Results evaluation, in particular for home-work trips, presented strong similarities against travel survey and census data on the Boston metropolitan area. Gundlegard et al. (2016) proposed a process for travel demand and route travel flow estimation as well as for mobility metrics extraction. Their analyses were based on CDR datasets provided from Ivory Coast and Senegal territories. In this work, derived O-D matrix and developed methods were not evaluated due to the lack of validation data.

Although CDR data have supported interesting findings for O-D extraction, their limited temporal granularity could introduce biases (on geographic and demographic levels) since the location of a mobile phone owner is recorded only when the user calls, sends a message or makes data connexion, which means that user's movements are unknown when he/she does not use his/her phone (White and Wells 2002; Calabrese et al. 2011a; Zhao et al. 2016). To alleviate this limitation, some published works combined CDR data with other urban transportation data sources like GPS data (e.g., from taxi ,private car or mobile phone application) (Huang et al. 2018; Widhalm et al. 2012; Hoteit et al. 2016), smart-card data (Huang et al. 2018), travel survey and existing transport model (Wismans et al. 2018). Meanwhile, (Fiadino et al. 2017) have shown that the situation has changed and that new CDR data quality has improved thanks to the high penetration of "always connected" terminals. Furthermore, Toole et al. (2015) have combined CDR data with route information from crowd sourced geospatial data, census and travel survey in order to impute missing attributes like transport mode or purpose and to more accurately estimate O-D matrix in three metropolitan cities. Results are promising even if some differences might be important.

Cellular network signalling data are another form of passive mobile phone traces collected from providers for technical management purposes. User location is recorded regularly in time, independent of whether the user is using his/her phone or not and until the phone is turned off. Therefore, mobile signalling data are more likely to be appropriate for O-D matrix estimation and show enormous potential for travel demand modelling since they capture all network-based events providing higher spatiotemporal granularity. Few existing works applied these data for traffic modelling. Fiadino et al. (2012) presented data-driven approach using signalling traffic of a 3G cellular network to extract vehicular trajectory patterns. Tettamanti et al. (2012) used 2G signalling data to estimate the route choice using traffic assignment macroscopic simulation and evaluated the method for one O-D pair in the test area. More recently, in a study conducted in the Paris region (Bonnel et al. 2015), authors used signalling data collected from 2G network in 2009 to produce O-D matrix of individual travels and compared them with the local household travel survey. They obtained similar estimations for O-D pairs with high traffic. Same form of data have been analysed in (Ni et al. 2018) to explore the impact of several factors such as population and transport accessibility on urban travel flow in Hangzhou (China). Huang et al (2018) proposed a data-driven real-time mobility model for the city of Shenzhen (China) that combines the advantages of 2G mobile phone signalling records (of one day) and urban transportation data. The model validation was performed by comparing the predicted mobility flows and the travel demands obtained from the same signalling data used to build the model, as no other data were available for evaluation. Hence, although they showed promising results, using the same data for modelling and validation may have an impact on the validation process. None of the previous works appears to have achieved reliable complete O-D matrices using only 2G signalling data.

Overall, there is a wide variety of literature available about techniques and methods to generate origin-destination flows using mobile phone location data. However, there are still several limitations to be considered in future studies. First, the majority of researches have focused on CDR data which are characterized by a low temporal resolution since they rely only on mobile phone communications as mentioned earlier. The impact of this data feature is not very well discussed in these studies and the corresponding results would be highly biased especially for those people who do not frequently use their mobile phones. Other researchers have investigated also the triangulated CDR data (e.g. in US cities) which consist of CDR records with estimated cell phone coordinates rather than cell tower location. It shall be noted that although these data have a higher spatial resolution, an additional complex pre-processing module is required in the collection infrastructure to estimate the mobile phone's coordinates based on a set of measurements (e.g. number of surrounding cell towers, signal strength). Hence, CDR data-based approaches (for both forms) could not support rigorous applications such as long-term and real-time dynamic O-D estimation. Moreover, few studies have addressed the validation of the outcome and the accuracy of results and some of them have used the same data for the matrix estimation and validation. Indeed, conventional validation methods are still missing to fully verify the consistency of estimations and should take into account that the validation or ground truth datasets attributes do not often match with those of massive data (Bonnel et al. 2018; Chen et al. 2016). Such evaluation is critical as mobile phone data are emerging as a new, promising data source for policy makers to guide transportation development and especially if it is envisaged to use the inferred knowledge for transport network optimisation or for planning and decision making purposes. Furthermore, representativeness of the analysed data is of prime importance when dealing with population mobility modelling in a territory. Existing works employ different types and volumes of mobile phone data but they rarely discuss their proper attributes, the effect of data processing and the actual representativeness of the findings. For instance, by exploring CDRs, only a partial amount of individuals' whereabouts is captured as the

location details perceived from CDRs are biased by the actual terminals' activity patterns. Accordingly, it is required to delve into the hidden characteristics and issues of these network operation–based data. Additionally, in numerous published studies, authors have evaluated only travel flow structure and trip distribution instead of trip volumes (Graells-Garrido and Saez-Trumper 2016; Wismans et al. 2018) since trip extrapolation to absolute level is not straightforward. Therefore, adequate methods to expand inferred O-D matrices still need to be investigated to exhibit the whole population.

The aim of this research is to advance the state-of-the-art on the potential of network-based signalling data for origin-destination flow matrix extraction. To that aim, our method explores a recent mobile network-based signalling dataset, collected in 2017 from both 2G and 3G cellular networks in the Lyon French region area (about 44,000 km^2). By using such dataset, we will intend to reduce the gap on the various cited limitations of using CDR data and 2G or 3G signalling data separately. Besides, we present a different convenient process to derive O-D matrix after pre-processing the raw data. We introduce new techniques in each step by considering the specific positioning information included in our signalling dataset. Our method includes a novel indicator-based filtering which helps to extract consistent mobility information from signalling traffic in addition to an extensive validation step using the conventional O-D matrix generated from a local travel survey. An application of the proposed method in a French region shows the capability of the by-product signalling logs to reproduce reliable trip flows in a large-scale area without introducing additional high-cost travel data. The following section describes in details the used dataset and the study area.

3. Data sets and study area description

3.1. Cellular Signalling Data

Mobile network data are continuously collected by telecom operators for billing and technical measurement purposes. Among mobile network technologies, we focus on the traditional GSM[2] network, which provides 2G services, and the UMTS[3] network for 3G ones. Both GSM and UMTS networks have different infrastructures, but they still work with the same coverage concept. Each antenna covers a cell, which belongs to a larger Location Area (LA)[4]. Typically, tens or even hundreds of cells share a single LA. In many studies on mobile networks, the theoretical coverage area of the cells is represented by means of Voronoi polygons.

In this paper, we present analyses of mobile phone signalling data collected from the cellular network of Orange. The explored dataset consists of 2G and 3G signalling records of over two million anonymous mobile phone users in June 2017. For legal privacy restrictions, data from only one day are used and include a total of about 300 million records of device transactions. Concerning the spatial dimension, this dataset covers the whole extent of the Rhône-Alpes region in France, thus allowing for estimating origin-destination flows within this territory. Fig. 1a presents the cellular network coverage within the Rhône-Alpes region and the aggregation in 3G Location Areas. There are about 2,230 cell towers in the study area and each cell tower may handle several antennas.

The employed data represent the signalling traffic transiting through the 2G and 3G networks. Records include all the events that are generated by mobile devices or by the network itself (Smoreda et al. 2013). 3G traffic captures more logs than 2G traffic as a result of the extra internet services it is able to monitor. Each record in the dataset includes: the anonymised user ID, the event type, the coordinates of the cell tower serving the mobile phone and the assigned timestamp. Different types of signalling events are then captured in the explored dataset including:

i) communication events (i.e., calls and SMS);

---

[2] Global System for Mobile communications

[3] Universal Mobile Telecommunication System

[4] A "Location Area" is a set of cells (antennas) that are grouped together to optimise signalling.

ii) itinerancy events: handovers (i.e., cell changes during a communication) and Location Area Updates (LAU) (i.e., when crossing a LA or in idle mode);
iii) attachment/detachment events ;
iv) data/internet connections.

The above-mentioned event types constitute a larger set than those traditionally observed with event-driven data (e.g. CDR), thus explaining the higher temporal granularity of the network-driven data leveraged in this study.

3.2. Travel Survey Data

The Rhône-Alpes region authorities have conducted a travel survey for the first time, at the level of the whole region, between 2012 and 2015 (called EDR 2015). 37,450 individuals, aged over 11 years, have been surveyed, and 143,000 trips have been identified. Data have been collected by phone interviews using a representative sample of the region population. The Rhône-Alpes region has a total of 5.2 million population aged over 11 years and covers an area of 43 700 km^2. The survey sample has been constructed using geographical stratified random sampling. The geographical stratification corresponds to a zoning system of 77 zones (denoted as EDR-sectors) for the whole region (Fig. 1b shows the 77 EDR-sectors (codes from 101 to 177) and, with different colours, their aggregation in 14 macro-zones). Each EDR-sector involves at least 450 surveyed individuals. The largest metropolitan area in the territory is the city of Lyon, which concentrates nearly 25% of the inhabitants of the region.

The survey collects socio-demographic characteristics of the individuals and of the household, as well as information about all the trips that were made the day before the survey (from 3:00am to 3:00am next day). The most important attributes characterizing a trip are: transport mode, start and end time of the trip at minute-level granularity, activity at the origin and activity at the destination, location of the origin and location of the destination. Data has been collected through three waves in 2012/2013; 2013/2014 and 2014/2015 from late autumn to early spring gathering only working day trips. Survey methodology is similar to other travel surveys conducted in urban areas in France (CERTU 2008).

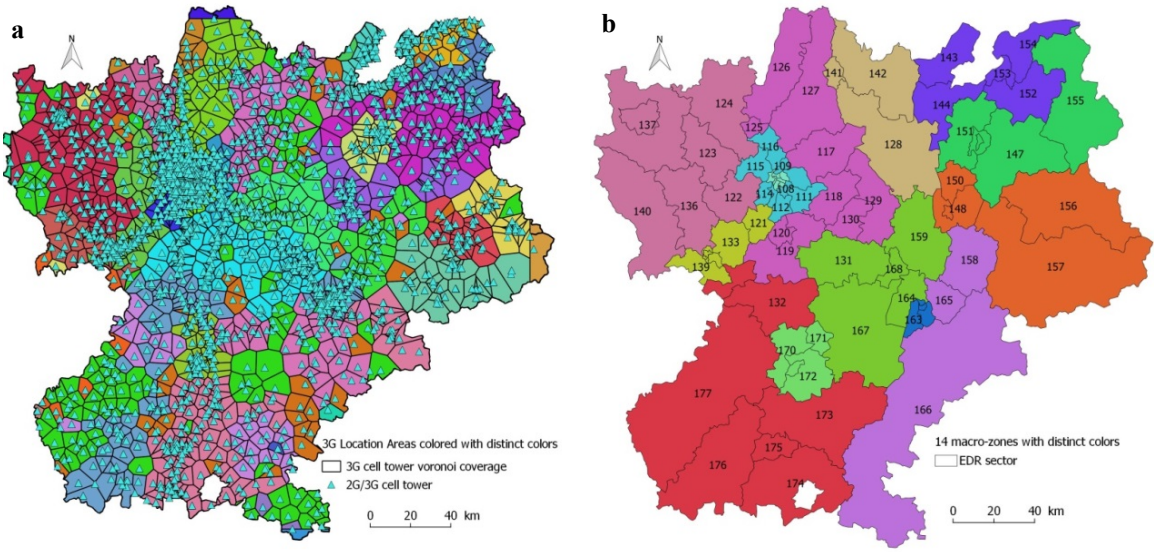

Fig. 1 (a) Cell tower distribution and cellular network coverage. (b) Aggregation of EDR sectors into 14-zone zoning system in the Rhône-Alpes region

4. Methodology

In previous work (Bonnel et al. 2018), we introduced a first simple approach to generate an O-D matrix from signalling data. In that approach, only 3G data were considered in the analyses. All observed users in the dataset have been involved to study travel flows and hence one generic expansion factor has been applied to the entire region to expand extracted trips. In this paper, a more complete and flexible workflow is presented in order to transform mobile phone signalling data into comprehensible O-D flow matrices (Fig. 2), supported by extensive validation step. The proposed workflow consists of :

i) identifying users' home locations;
ii) analysing the cell phone activity indicators to better characterize and understand the dataset;
iii) filtering the detected residents based on their activity indicators to only retain users whose device traces appear pertinent to study their typical daily trips;
iv) extracting and scaling up trips according to estimated expansion factors to aggregate them at the travel survey zoning level (EDR-sectors) and infer the O-D matrix.

It is worth to remark that our approach solely leverages cellular signalling data for user filtering and trip extraction, while it depends on travel survey and census data in relation to zoning (for spatial aggregation) and trip scaling (for determining the set of expansion factors), respectively.

Due to user's privacy protection regulations, mobile phone data analyses are only allowed within a study period of maximum 24 hours. Hence, we have analysed the data of June 1st, 2017. It is a working day –Thursday-, which is traditionally considered (in transportation surveys) as representative of an average weekday. In order to be comparable to EDR, cellular network-based data are collected from 1st June 3:00 am to 2nd June 2017 3:00 am.

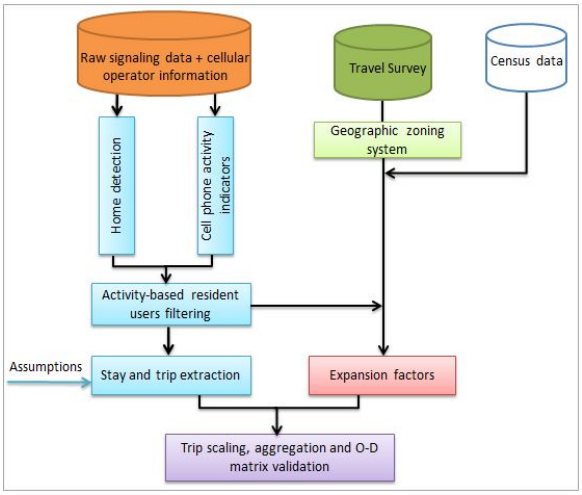

Fig. 2 Workflow of Origin-Destination matrix construction from cellular network signalling data

4.1. Home Location Detection

Cellular network-based data do not contain any socio-demographic information that characterizes the users due to privacy concerns. However, a plethora of works have investigated basic mining solutions to detect users' home location (Calabrese et al. 2011b; Ricciato et al. 2016) in combination with home activity identification (Frias-Martinez et al. 2010; Jiang et al. 2017).

In our study, the focus is on the detection of users' home locations by considering only those that reside in the region of interest and expanding the obtained estimations to the whole region based on population census data. The adopted method to compute home location consists of the following steps:

1. Filter user traces to select only those occurring at night time from 3:00am to 7:00am and from 10:00pm to 3:00am;
2. Filter user traces to keep only device events that could be generated in a stationary state such as (video) Call, SMS, Attachment, Detachment, Data, and periodical events (e.g. LAPU);
3. For each user, extract all observed cell towers to which the user's cell phone has been connected;
4. For each user, derive the most frequent observed cell tower, assign it to the corresponding sector and consider it as the home location zone of the user.

In order to detect home location, the study period has to be divided between day and night time windows. In the first step, day period records have been excluded. This time allocation choice takes into account the mobility behaviour of the involved users in the dataset. Fig. 3a and 3b show the temporal distribution, during the observed period, of LAU and LAPU events, which can be reasonably considered as representative of the mobile/stationary and active/idle behaviours, respectively. Both distributions show the typical life-cycle of individual movement with high observed mobility (frequent LAU) approximately between 7am and 10pm and less active, therefore more stationary, devices during night time (high LAPU). Also, the presented method differs from the existing home detection methods by the inclusion of an event filtering step (step 2) that considers the existence of more kinds of events than those that can be found in traditional CDR datasets, thus excluding mobility-related events such as handover and LAU. Hence, a new approach is introduced to adapt existing algorithms to signalling data. By applying this method, 1.27 million resident users are identified in our mobile phone dataset related to the Rhone-Alpes region. This corresponds to 62% of all mobile phone users that are observed in our dataset, and about 25% of the total region population.

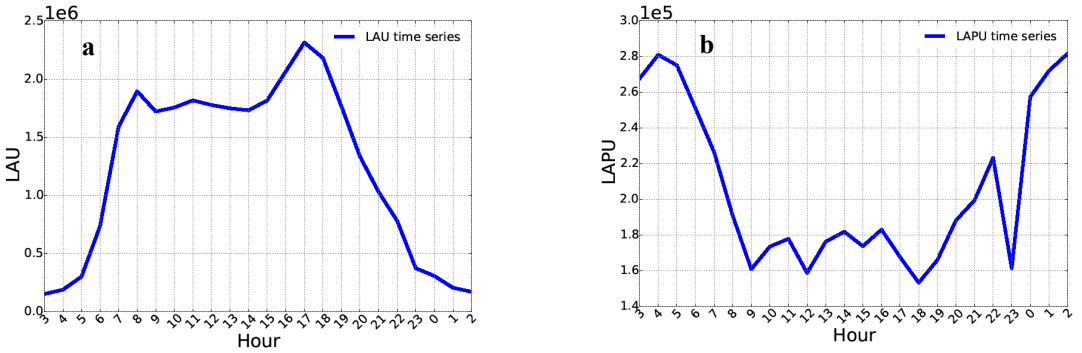

Fig. 3 Temporal distribution of signalling events (a) Location Area Update (LAU) and (b) Location Area Periodic Update (LAPU)

4.2. Cell Phone Activity Indicators

Given that mobile phone data describe the communication traffic monitored by cellular networks, it is worth to highlight that such data may not always reveal expected individual behaviours (Wang and Chen 2018). Particularly, the signalling data consist of a massive amount of record streams that are passively generated for operation and technical purposes by the telecom provider. For that reason, it is of utmost importance to leverage the hidden data attributes and understand the potential intrinsic anomalies (e.g. noise) to properly conduct a signalling data-based mobility research and avoid estimation errors. Accordingly, relevant mobile phone activity indicators are introduced in this study in order to measure the amount of logs per user, to examine the uniformity of traces distribution over the study period, and to identify the outlier devices, which should not be included in the O-D flow estimation process. In the following, we make the assumption that each mobile phone (terminal) corresponds to one user.

- *Number of observations (NO):*
  This indicator measures the number of records (logs) for each terminal over the observed day. Fig. 4a shows that records frequency on the dataset widely varies among observed devices. Specifically, in relation to our case study, around 99% of users have less than 450 events (which is considered as a high value comparing to CDR data) and about 10% have less than 7 records. It is possible to consider a variety of reasons why some users are rarely observed. For instance, users might be travelling on the specific observed day, thus either leaving the region at early morning or entering the region at late night. This is especially true in our case study of the Rhone-Alpes region as it is characterized by an important number of major transit station areas (e.g. regional transportation hubs such as airports, important train stations, major highways, etc…). Besides, terminals that are turned off or left inactive for a long time (e.g. during the observed day period) should naturally generate only a few records. A small amount of devices (1%) seems to be extremely active with a very high number of observations (more than 1,000), which is not imputable to a typical human behaviour, but very likely caused by device anomalies (e.g., buggy terminals continuously sending messages).

- *Average Inter-event Time (AIT):*
  This measurement is largely used when dealing with individual temporal data. It gives an overview of the average time between users' successive observations. Fig. 4b shows that AIT values range from 1 minute to 1372 minutes (22 hours), and the average value is about 40 minutes (while in previously studied datasets the average time would typically be longer than four hours (González et al. 2008; Calabrese et al. 2011b)). The large majority of users (99%) are characterized by an AIT lower than 200 minutes. Obviously, those terminals (3%) with AIT smaller than 1 minute (few seconds) represent devices generating a lot of records.

- *Maximum Inter-event Time (MIT):*
  Since with AIT the entire 24 hours are covered, this could have an impact on its values given that, during night-time, devices are typically less active than the rest of the day. Therefore, in order to select the users for our study, we propose to examine the maximum inter-event time during an interval of time (7:00am-10:00pm) that excludes deep night and early morning. The MIT distribution is more skewed to the right (Fig. 3c), showing that 70% of users present a MIT lower than 180 minutes. In particular, Fig. 4c shows a steep increase on 180 minutes that corresponds to the typical time period (3 hours) at which the mobile phone automatically generates Location Area Periodic Update (LAPU) during idle (inactive) mode. Consequently, the remaining observed users (30%) having MIT larger than 180 minutes are either not present in the study area during the whole [7:00am-10:00pm] time window or were disconnected from the network (e.g. mobile phone switched off) for a certain time longer than 3 hours.

- *Entropy (H):*
  This metric consists in measuring the uniformity of the number of signalling events per user over the 24 hours. It gives more precise information about the temporal distribution.
  Entropy is defined as: $H(X) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i))$. For our case, we consider X as the distribution of the records of a user over 24 hours and $p(x_i)$ as the fraction of the records in the 1-hour time slot x_i. Entropy values were normalised to get value range between 0 and 1. Fig. 4d shows that about 5% of the devices have all observed traces in only one hour time-slot, which is described by an entropy value of 0. While 99% of devices have an entropy value lower than 0.9 (more uniform behaviour). Those mobile phones with entropy close to 1 have extremely uniform generated records that could depict machine behaviour.

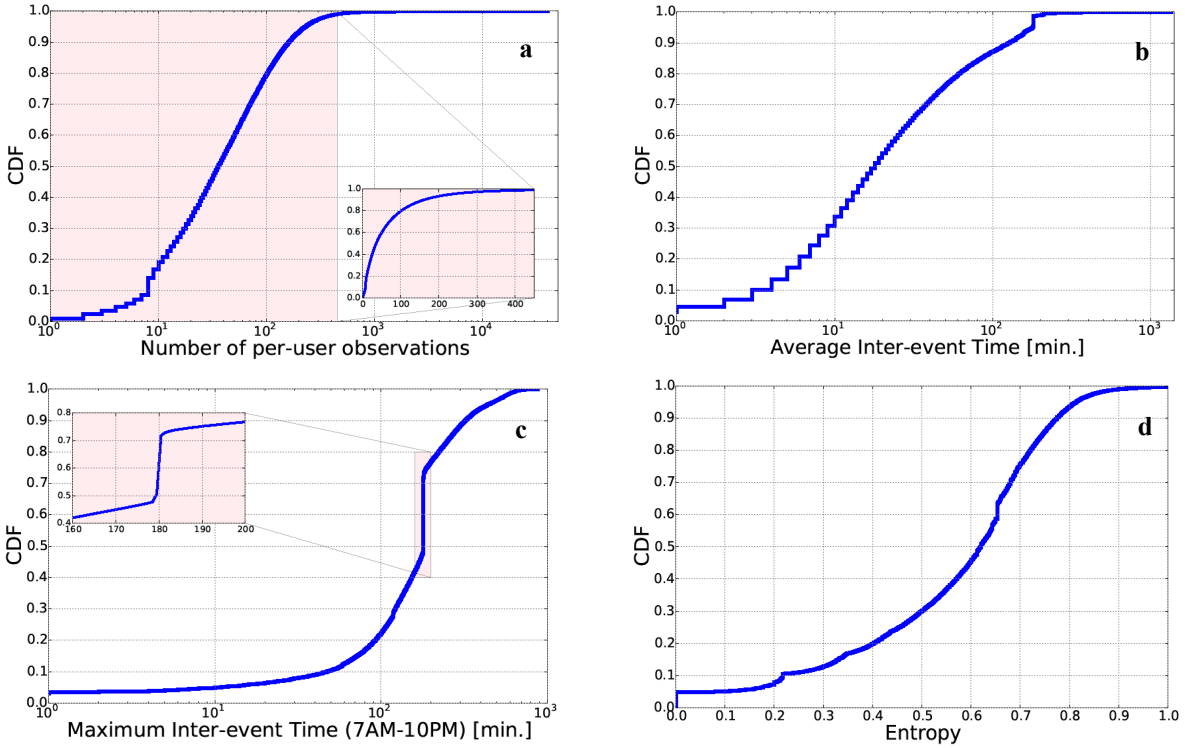

Fig. 4 Cumulative distribution function (CDF) per user of (a) number of observations (b) Average Inter-event Time (c) Maximum Inter-event Time [7am-10pm] and (d) Entropy

4.3. User Filtering based on Cell Phone Activity Indicators

As previously mentioned, processing mobile signalling data without proper understanding of their content can lead to important estimation errors. For instance, due to increasing pervasiveness and wide adoption of embedded connected devices, 2G and 3G telecom networks do not only capture human mobile phone communications, but also transactions from devices and sensors that use the same technology (e.g., Internet of Things) for different purposes (e.g., performing batch operations, continually collecting and uploading data to servers, etc.).
Thus, before using the signalling data, it is fundamental to clean the data in order to properly select human-related tracks. To this end, we propose to leverage the cell phone activity indicators introduced in Section 4.2 to further filter the retrieved set of resident users (i.e., those for whom home location was detected (Section 4.1)) as not all of them are necessarily captured and/or have sufficient observations during the study period. Our filtering approach requires the definition of thresholds associated to the indicators' values and consists in a pipeline of selection rules. Based on the explanations given in Section 4.2, a rationale for identifying each rule is presented below:

- *Maximum Inter-event Time (MIT) ≤ 180 minutes:* the studied original dataset include all traces of observed mobile phone users in the Rhone-Alpes region over a 24-hour period. For multiple reasons such as travelling, there might be a fraction of users who are only observed in few hourly slots (e.g. one hour). Since the identification of residents is performed based on night period (Section 4.1), it is required to check whether these user residents are observed during the day period to be able to extract their typical mobility patterns within the region. This allows also to keep common daily behaviours. According to the network system if a mobile phone remains inactive during 3 hours, a periodic event LAPU is

generated (as explained in Section 4.2) to maintain its location updated within the cellular infrastructure. Theoretically speaking, it follows that each user should have at least 8 observations if his mobile phone is switched on during the 24 hours. We consider this network-based measure (i.e. 3 hours) to set up the maximum time allowed between consecutive observations in order to ensure the presence of the detected residents in the region during the day period;

- *Entropy (H)* ≤ 0.9: as stated before, newly collected signalling data involves also Machine-to-Machine communications between objects equipped with SIM cards, which should not be considered in our analyses. In the dataset, 1% of devices generate a very high number of records and according to entropy distribution 1% of devices have an entropy higher than 0.9. Thus, devices with entropy higher than 0.9 are filtered out as they are extremely active and have too much uniformity in their observations' distribution during the 24-hour period. This fraction of devices are not considered to be handled by individuals and hence do not reflect regular human mobility patterns;

- *Number of observations (NO)* ≥ 4: this rule is set according to the stationary activity (i.e. stay location) and trip definitions. A necessary condition to detect a stationary activity location from signalling data is the time spent in that location (i.e. the duration of consecutive observations in the same zone). Consequently, to identify that an activity has undertaken, at least two observations are needed. Since two activity locations (origin and destination) are required to determine a trip, 4 observations at least must be recorded for each user. This condition is set to ensure the minimal requirements of trip detection. Section 4.4 explains in detail the activity and trip definitions used in this study.

The combination of these indicator-based rules implies indirectly the exclusion of detected residents who are not captured in day period (e.g. they might be visitors) and whose travel patterns cannot be representative. After the filtering process, a large sample of 985,483 users is still retained. This represents approximately 77.3% of the total users for whom a home location could be attributed, and around 50% of observed users in the original dataset. The resulting distributions and average values of the different cell phone activity indicators for the selected resident users are reported respectively in Fig. 5 and Table 1.

It is worth to note here that we apply a "soft" filtering process that aims to guarantee the representativeness of the detected residents and to filter out devices that are not suitable to study the individual travel behaviour. Indeed, our main concern is to keep the sample as large as possible to ensure the relevance of results. We also remark that it is not appropriate to perform a very restrictive filtering to keep only (highly) active users, as done in (Alexander et al. 2015; Fiadino et al. 2017), since that could affect the representativeness of the users' sample (e.g. overestimation of frequent users) and could lead to estimation biases.

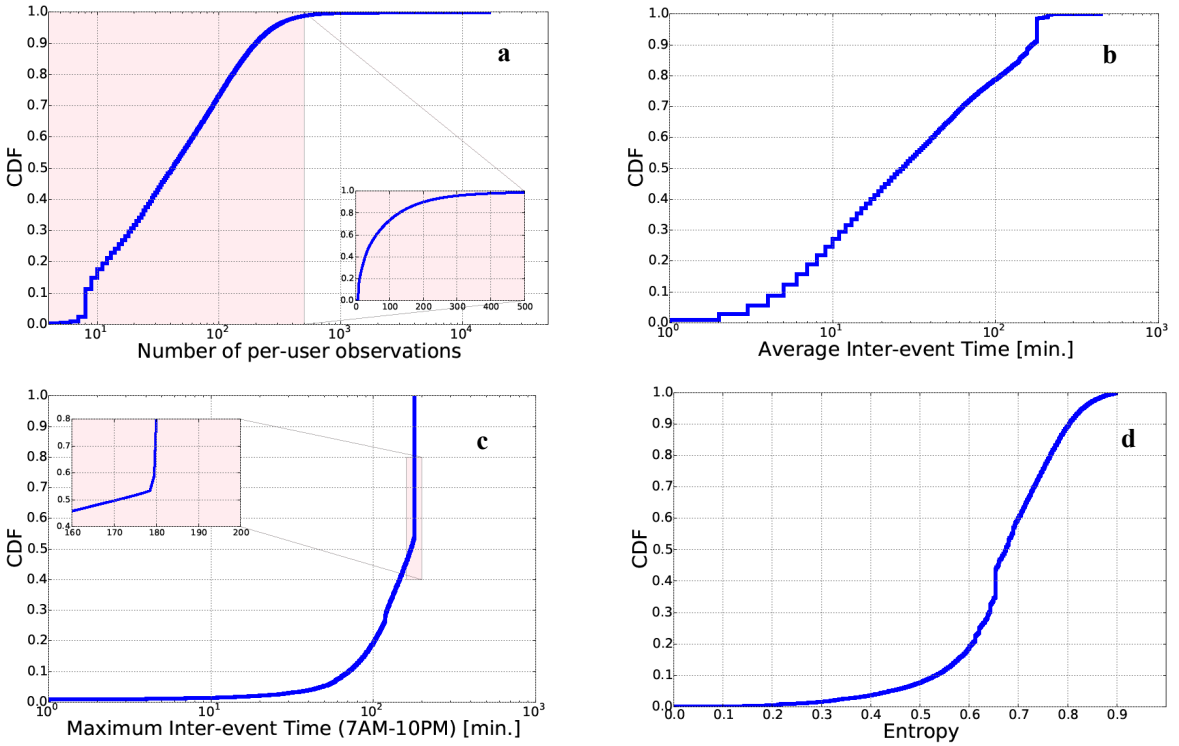

Fig. 5 Cumulative distribution function (CDF) per user of (a) number of observation (b) Average Inter-event Time (c) Maximum Inter-event Time [7am-10pm] and (d) Entropy after filtering process

Table 1 Average cell phone activity indicator values for filtered resident users

Indicator	**Average value**
Maximum Inter-event Time [7am-10pm] (MIT) in min	143.12
Entropy (H)	0.67
Number of observations (NO)	85
Average Inter-event Time (AIT) in min	54.38

4.4. Trip Detection

After identifying and filtering the resident users who are potentially appropriate to study the origin-destination matrices, trips can be finally extracted. Recently, some researchers (e.g. Chen et al. 2016) have raised the issue of the unclear trip definition in various mobile phone data-based studies. Thereby, we aim in this paper to use an appropriate definition of the trip which is applicable to signalling data and, at the same time, coherent with what has traditionally been used in travel surveys, towards a meaningful and relevant comparison of results.

A trip has been defined by CERTU (the French agency for transport network and urban planning) for the purposes of the EDR as follows (CERTU 2008): a "*trip is the movement of one person conducted for a certain purpose on a transport infrastructure open to the public, between an origin and a destination with a departure time and an arrival time using one or more means of transport*". Hence, to apply this definition for trip extraction, it is necessary to identify an origin and a destination and therefore a stationary activity in both locations.

With the huge amounts of footprints and high spatiotemporal resolution, signalling data collected from mobile devices provides an unprecedented scale of observation. These proper characteristics allow quantifying user's trips at a higher

level of geographical detail (e.g. cell area) for which travel surveys cannot provide accurate estimations. Since the scope of this paper is to generate an O-D matrix and to be able to validate it at the same level for which EDR data are available, the trip extraction method is presented in the following at the EDR-sector level. The detailed experiments and additional data processing steps needed to study signalling data at a more fine-grained spatial level will be matter of future work.

To extract trips, stationary activities need to be identified first. Thus, consecutive observations of a user in EDR-sector zone within a minimum stationary time threshold are considered. However, the size of the zones (average area of EDR-sector is 582 km²) and the fact that the user is travelling should be taken into account. In case of large areas, consecutive observations might be in the same zone even while the user is traveling: this grounds some lower bounds on the time threshold that can be applied. Therefore an activity assumption has been defined as follows: if an individual is present for at least a given time threshold in a sector, she/he performed a stationary activity there and the origin or the destination of a trip is located in that sector (the choice of the time threshold and its impact are discussed in the result section).

Based on the previous hypotheses, the following pipeline is proposed to identify users' trips:

For each user *i* :

- Extract all the observed location points and associate to each location an EDR sector with the help of a conversion table.

  Cell tower → Sector

- Sort the sequence of extracted locations by timestamp , denoted by: $S_i=\{s_i(1), s_i(2),...,s_i(n)\}$, where $s_i(k) = (t(k), l(k))$ for $k = 1,...,n$, and $t(k)$ and $l(k)$ are the time and location of the k^{th} observation.
- Extract only locations for which the EDR sector is consecutively the same for a time duration t,

  where $t \geq threshold_{min}$ (minimum stationary time): we obtain activity locations.

Trips are then evaluated as paths between user's activity locations at sector level. Each *trip (U, O, D)* is characterized by user id *U*, origin location *O* and destination *D*.

4.5. Expansion Factors Definition

Albeit large, the analysed mobile phone user sample does not represent the full population. Therefore, extracted trips need to be properly scaled in order to be representative of the mobility of the full population. After performing home detection, an expansion factor can be applied for each filtered user as the ratio of the census population and the number of residents estimated by the cellular signalling data in his home sector. It follows that residents with the same home sector have the same scaling factor. Accordingly, all resident users living in a particular home sector are equally weighted. The expansion factors are therefore calculated on a sector-basis as in equation (1): for all mobile phone users whose home is detected in a given sector s_i, the expansion factor corresponds to the ratio between the sector's population (from census data) and the number of resident users identified in that sector. Hence, each trip of a given resident user is scaled according to the same expansion factor of his home sector.

$$F_{exp}(s_i) = \frac{Population\ of\ s_i\ (over\ 11\ years)}{Nb\ of\ home\ locations\ detected\ in\ s_i} \quad \text{where } \textit{nb of home locations} \text{ means number of resident users} \qquad (1)$$

Fig. 6a and 6b illustrate the probability distribution of the expansion factors through sectors before and after the resident user filtering step. The 1st, 2nd and 3rd quartiles of the expansion factors after filtering are 4.32, 5.39 and 6.71 respectively.

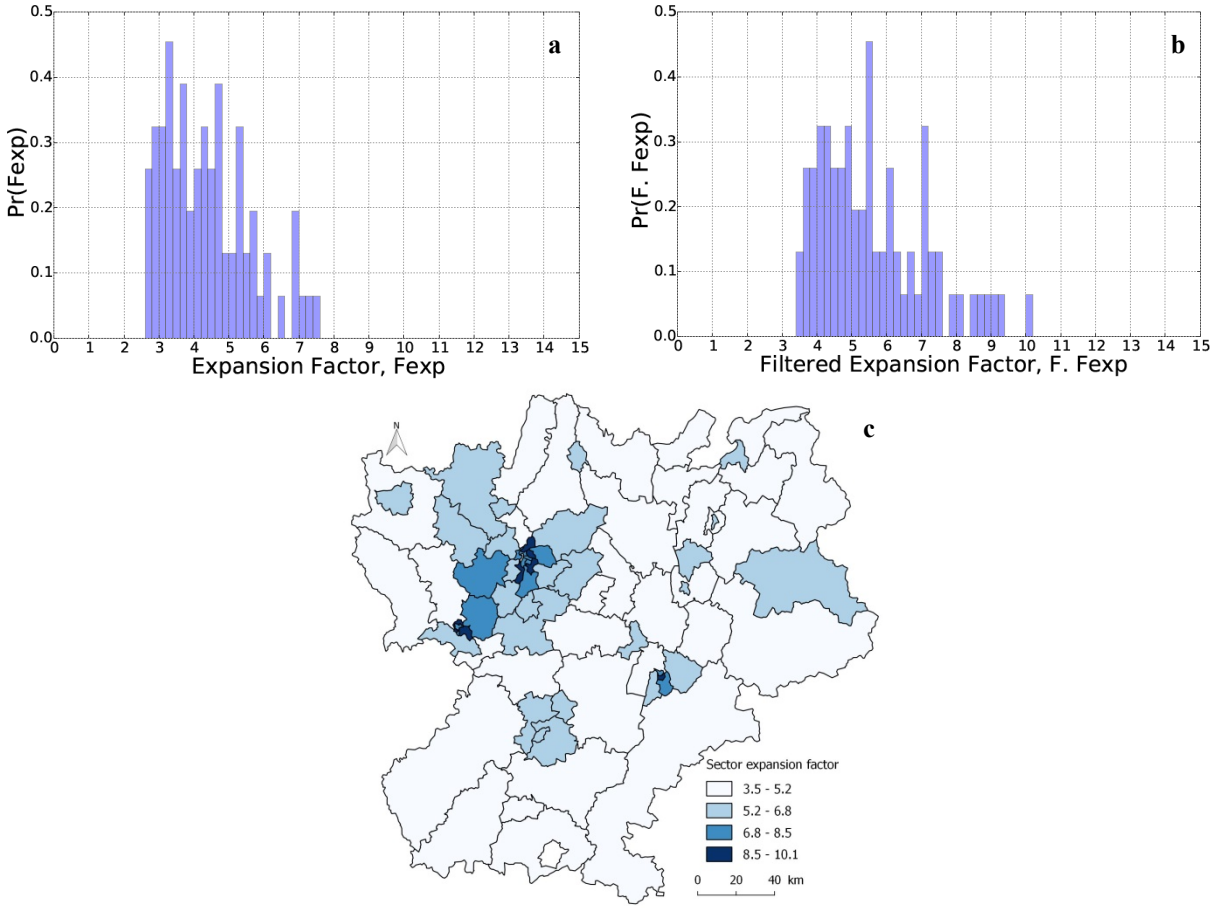

Fig. 6 Probability distribution of sector expansion factors (a) before and (b) after filtering (c) the spatial distribution of expansion factors after user filtering

The spatial distribution of home expansion factors (Fig. 6c) shows that the sectors in the metropolitan regions of the study area tend to be more heavily weighted. One of the potential reasons is that actually in these areas the number of subscribers using the 4G cellular network –not covered by our study- is expected to be higher than in the other zones. Thus, a lower fraction is observed in the available dataset (i.e. 2G and 3G data), which yields a larger expansion factor.

5. Results and validation

After identifying the users' home, filtering the residents, extracting and expanding the trips, the origin-destination matrix can be constructed over all the 24-hour period. As stated in Section 4, the definition of a trip leads us to the assumption of the minimum activity stationary time. Considering the size of a sector, most trips between two zones are made by motorized transport mode, except for pairs of adjacent zones. According to EDR data, the average duration of a trip to cross a sector with motorized mode is estimated to be lower than one hour. Also, the fact that consecutive observations might be in the same sector even while the user is traveling grounds some lower bounds on the time threshold. Meanwhile, the sampling rate of events in mobile phone footprints should be considered. As stated in section 4.3, the measured average inter-event (AIT) time after user filtering was about 54 min for the 24-hour period and about 50% of the users has AIT more than 30min. Also, the AIT is calculated for only day period [7am-10pm] as

the AIT measured on the basis of 24 hours could be affected by the long night period inactivity of mobile phones. It was found that half of users have AIT of the day period [7am-10pm] less than 19min. Therefore, it was decided to apply different stationary time thresholds to test the impact of such parameter on the number of generated trips, as the produced O-D matrix elements are expected to change according to this threshold. Hence, thresholds are tested between 30 minutes and 60 minutes to show the sensitivity of the trip estimation at different levels. It does not appear recommendable to consider time thresholds which are lower than 30 minutes, as multiple false-positive stationary detections may occur, yielding false-positive trips.

5.1. Trip frequency distribution

In this section, we present the distribution of the number of trips on a typical weekday with respect to the users before expansion. The idea is to study how this distribution behaves without the additional assumption of scaling, as the latter could impact the trip distribution on individual and spatial level.

The frequency of total trips per user for two stationary thresholds (30min and 60min) is shown in Fig. 7. The two distributions have a long tail, with first, second, and third quartiles of 1, 2 and 3 trips per user per day, respectively, demonstrating that the large majority of users have a reasonable small number of trips. As expected, a higher threshold of 60 minutes tends to give a lower number of trips (the threshold impact will be analysed in more details in Sections 5.2 and 5.3).

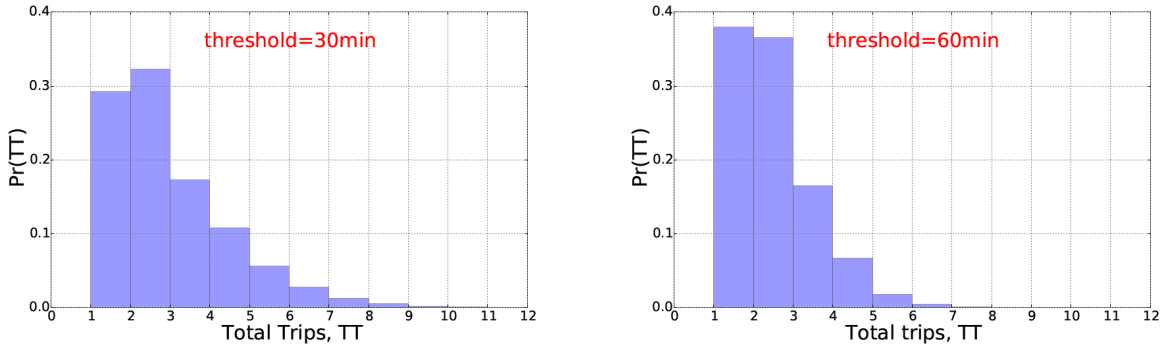

Fig. 7 Probability distribution of total trips per user, before expansion, with a stationary threshold of 30min and 60min

5.2. Origin-Destination matrix at sector level

The overall shape of the origin-destination matrix inferred from signalling data at EDR-sector level has been investigated and compared with the travel survey-based matrix.

The data from the EDR contain all the trips reported by residents of the Rhône-Alpes region irrespectively of the purpose and the duration of the activity collected on working days. However, the assumption of the minimum stationary time in a sector has been considered in order to identify an origin or a destination in the case of mobile phone data. Therefore, we do extract information from the EDR data and apply time thresholds to avoid considering false activities and therefore false trips when dealing with the comparison.

To have an overview of the generated O-D flows based on the mobile phone sample, we considered the number of O-D pairs for which a number of trips is estimated in the cases of EDR and mobile phone signalling matrices, respectively (note that the total number of possible O-D pairs is 5,929). In Fig. 8, we observe that in the travel survey matrix, less than the half (about 40%) of O-D sector pairs are assigned to trips, while in the mobile phone-based matrix, we obtain a yield of 95%, for all thresholds that were considered. This confirms the sampling bias that is inevitably present in O-D matrices that are constructed based on travel surveys. Indeed, it is cost prohibitive to obtain, via surveying, sufficient observations to produce an O-D matrix at reasonable level of geographic detail. On the other hand, it is relatively cheap to get a large sample from cellular network-based data, which reduces the zero-cell problem for such geographical level, and which enables the investigation at a higher geographical level.

Moreover, signalling data can more easily cover large-scale geographic areas as the collection is not dedicated but rather an operational by-product.

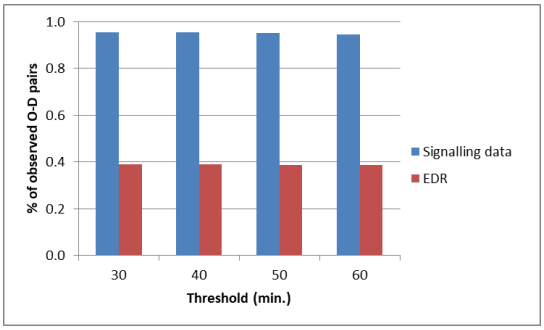

Fig. 8 Distribution of number of observed O-D sector pairs according to the stationary threshold

5.3. Origin-Destination flow matrices comparison

In this study, the aim is to test the potential of network signalling data to infer reliable origin-destination matrices and to investigate similarities and differences of the results with the traditional survey estimations. Therefore analysis is performed to compare both the structure and flows of O-D matrices from the two data sources.

While travel survey data can be representative of the population at sector level, combining origins and destinations typically leads to the fact that they are not representative anymore since the number of observed O-D pairs is too small , as explained in Section 5.2 and shown in Fig. 8. The confidence intervals are very wide for many O-D pairs. Therefore, the 77 EDR sectors are aggregated into 14 macro zones (Fig. 1b) in order to produce a relevant origin-destination matrix, which gives a sufficient number of trips for most of the origin-destination pairs in the EDR matrix. This enables a comparison with the mobile phone data matrix, which has also been aggregated to correspond to the 14-zone zoning system. The analyses are presented regarding the correlation between the two matrices after expansion and at macro-zone level by removing the intra-zone pairs since the focus here is on inter-zone flows. Table 2 summarizes the total number of trips from signalling data and the EDR.

Table 2 Total number of inter –zone trips from signalling data and the EDR (aggregation into 14 macro-zones)

Stationary activity time threshold	60 minutes	50 minutes	40 minutes	30 minutes
EDR (in thousand)	2,211	2,260	2,344	2,448
Mobile phones (in thousand)	1,607	1,743	1,905	2,108

The amounts of trips from the two sources are much closer when a stationary time around 30 and 40 minutes is considered. With such an interval, the majority of the sectors can be crossed by travellers: these thresholds identify activities which do not have short durations, but, on the other hand, they are still large enough to limit the number of false-positive trips due to excessively high travel time between sectors. Therefore, in the following analyses, we retain these values for activity threshold.

In order to highlight the weight of each O-D pair in the cellular data and travel survey-based matrices, the structure of the both matrices is visually compared in Fig. 9.

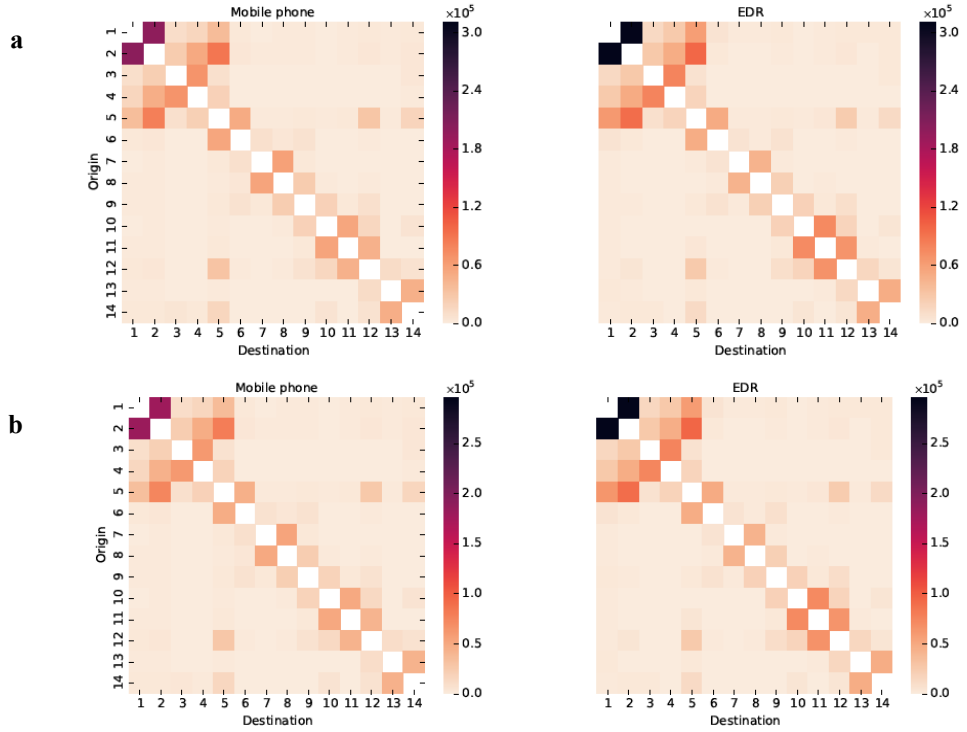

Fig. 9 Distribution of signalling data and EDR trips over the 14 macro-zones of Rhône-Alpes region for (a) 30min and (b) 40min

According to the distribution of O-D trips from mobile phone and EDR data with a stationary time of 30 and 40 minutes (Fig. 9a and 9b), the two flow matrices show very similar shapes even though the total numbers of trips are different. To see how different our results are with respect to the EDR, the Spearman's rank correlation is calculated at macro-zone level, and the result is $\rho = 0.95$ ($p < 0.0001$) for both thresholds. Hence, although both signalling data and survey-based matrices are developed using different techniques and technologies, they appear to resemble well.

We further investigate our results by means of a regression analysis aimed at supporting the comparison of the amount of flows corresponding to each O-D pair. That helps us to identify a coefficient of proportionality between the numbers of trips in each cell of the two matrices after the scaling step.
In addition to the total amount of trips, the coefficient of determination R^2 with value *0.96* between macro-zone trips gives a high-level indication that the distributions of O-D flows are similar with the following regression equations:

$$y_{ij} = 0.70 \times x_{ij} + 2{,}193, R^2 = 0.96$$ (Fig. 10a, related to a 30-minutes threshold),

$$y_{ij} = 0.66 \times x_{ij} + 1{,}964, R^2 = 0.96$$ (Fig. 10c, related to a 40-minutes threshold)

where y_{ij} is the number of trips from signalling data for the O-D pair ij and x_{ij} is the number of trips from EDR**.** Clearly, using large aggregation zones has a significant impact on correlation and results in a notable improvement in accuracy due to the reduction of sampling bias as a result of the aggregation. Results are much more satisfactory than the first studied approach (Bonnel et al. 2018) where we obtained $R^2 = 0.87$ for the same thresholds at macro-zone level.

As visually reported in the regression plot, two O-D pairs (see the two right-most points in Fig. 10a and 10c) have very high number of trips in comparison to all other O-D pairs. They correspond to flows between the Lyon conurbation (zone 1) and its suburban area (zone 2); the greater metropolitan area in the Rhone-Alpes region (see section 3.2). That is also shown in Fig. 9 for the O-D pairs 1-2 and 2-1. These two O-D pairs could have a strong

effect on the slope of the regression line. Therefore, a second regression analysis has been performed without considering the O-D flows between zones 1 and 2 in order to check the impact of such outlier on the regression results (Fig. 10b and 10d).

The regression line of the new model fits well for most of the O-Ds and provides slightly better parameters with the following regression equations:

$$y_{ij} = 0.85 \times x_{ij} + 877,\ R^2 = 0.95 \text{ (Fig. 10b, related to a 30-minutes threshold)},$$

$$y_{ij} = 0.80 \times x_{ij} + 788,\ R^2 = 0.95 \text{ (Fig. 10d, related to a 40-minutes threshold)}.$$

The slope is closer to one (0.85 and 0.8), and the constant (877 and 788) is relatively small compared to the mean number of observed trips on the O-D pairs (11,500) and the constant of the first regression (2,193 and 1,964). According to R2 value, 95% of the variance is explained by the fitted model. This means that, the majority of the O-D pair flows over the region match well irrespective of the travel demand volume.

Overall, the obtained Spearman coefficients and regression models show very encouraging results by the high-level of correlation between mobile phone and travel survey-based matrices in terms of both structure and volumes. This strong correlation is significant, given that users' trips were expanded based only on their home sector. Thus, these outcomes illustrate that the applied methodology based on the proposed pipeline of home detection, user filtering and expansion process could serve as a proper tool to extract accurate travel patterns from cellular signalling data passively collected over a limited period of observation (e.g., 24-hour period in our case). Moreover, results prove that the model is robust and could be applied throughout different socio-demographic profile areas; a crucial aspect when dealing with large-scale demand modelling.

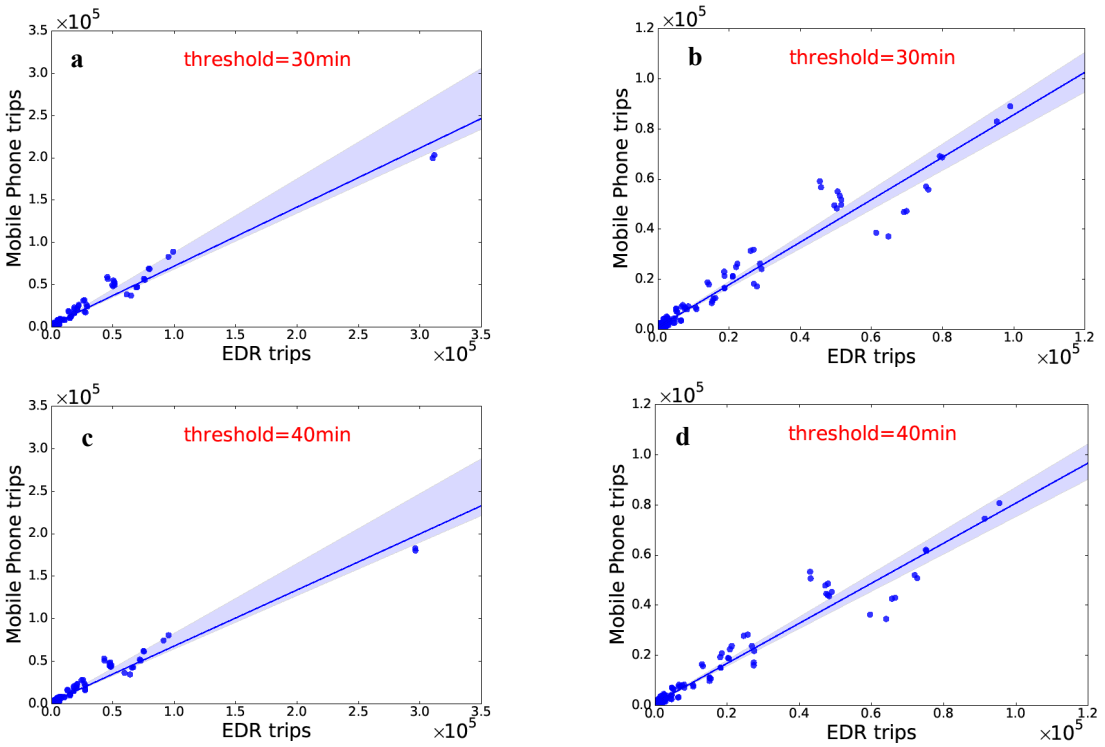

Fig. 10 Regression plots between the two matrices from signalling and EDR data with (a,c) all inter-sector pairs and (b,d) with all inter-sector pairs except Lyon O-D pairs

In addition to the previous analyses, the O-D matrices were explored to estimate, for each O-D pair, the percentage disparity between mobile phone counts and those from EDR to investigate the distribution of flow differences within the region. As a result, some of these percentages were very high and represent under-estimation cases with regard to EDR data. In most cases, these correspond to low (or very low) flows (less than 200 trips for EDR), and they mainly concern non-adjacent zones with similar percentage differences for both directions (4-6, 6-4, 12-7, 7-12, 7-10, 10-7, 3-9, 9-3). Also some remarkable flows of long distance trips between dense zones and suburban/rural zones seem to have more consistent estimate with signalling data. For instance, Fig. 11a and 11b show percentage differences of trip flows to and from Lyon suburban area (macro-zone 2) , respectively, obtained from the two data sources. It is clear that inflows and outflows of zone 2 towards distant zones such as 7, 8, 9, and 10 are significantly underreported (blue flow lines) in the survey-based data as more than 100% of EDR flow volumes are estimated with cellular data. In these cases, mobile phone data generate higher flows, which illustrates that travel surveys may not reliably estimate trips due to the representativeness of surveyed people sample and the sampling coverage.
However, the under-estimation cases with regard to signalling data mainly concern very high trip volumes (more than 60 thousand trips for EDR) between high population density zones, such as between the Lyon conurbation and its suburban areas. We suggest that this is caused by the minimum stationary time assumption, as a threshold of more than 30 minutes seems to be extremely large for those small sectors of the metropolitan area (e.g., the average sector area of Lyon city is about 9km^2). Thus, more investigation is required on this parameter which depends subsequently on the geographical stratification.

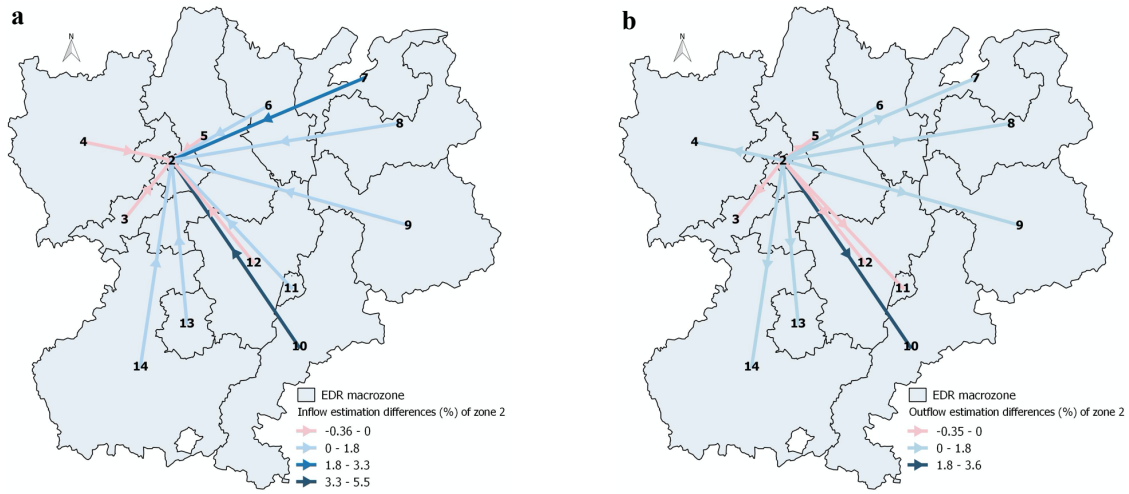

Fig. 11 Percentage difference of trip volumes estimated from signalling data and EDR data of the (a) inflows and (b) outflows of Lyon suburban area (macro-zone 2)

6. Discussions

For decades, traditional approaches such as travel surveys have been the major source of information for transportation planners to estimate origin-destination flows necessary for calibration and simulation of transport models. These travel surveys, although providing rich demographic details about the respondent and his/her trips, suffer from several drawbacks such as estimation bias due to the limited sample size of involved individuals, the high deployment costs and, subsequently, the low frequency of the gathered information making them rapidly out-dated and inappropriate for dynamic travel behaviour studies.

This study has demonstrated the feasibility of a large-scale Origin-Destination survey based on alternative passive real-world data. It depicts a first step towards building a complete and convenient framework to leverage rich cellular signalling data for individual mobility and travel flow modelling purposes.

In this direction, it appears however worth to report the major differences between conventional travel surveys and these emerging massive data as well as the key advantages of this study regarding previous work to fully clarify the potentials and challenges of the proposed approach. A first difference to highlight is related to the underlying population. The travel surveys normally consider only residents during the data collection process and select a small sample often representing much less than 1% of the whole population (Stopher and Greaves 2007) (e.g. in the EDR survey , the sample size is about 0.7%). Conversely, thanks to their higher penetration rate, mobile phone signalling data offer not only the possibility to involve significantly larger sample size but also to represent individuals from different groups (e.g. residents, visitors) who are present in the territory at a given time period. Hence, cellular data can help discover more relevant human travel patterns as well as the inter-individual variation considering persons with different profiles. In this work, we have focused on residents' behaviour to extract a typical O-D matrix. In particular, we have shown that due to the sampling bias, the travel survey-based matrix clearly suffers from the zero-cell problem as less than the half (about 40%) of O-D pairs are captured at sector level, while in the signalling-based matrix, we obtain a yield of 95%. Even after aggregation, signalling data depict significantly higher trip counts for those O-D pairs with very low flows in low population density areas (e.g. tens to few hundreds of trips) or those missing O-Ds from EDR. Additionally, EDR data appears to underreport flows for those long distance trips between dense zones and suburban/rural zones. In these cases, we expect that signalling data could produce more consistent estimations, since in large-scale areas, individual movements could be successfully observed and the threshold activity assumption seems to have very marginal effects on those long-distance flow estimations. Besides, datasets used in previous works, such as CDR data, do not capture the same population as in signalling data. As CDRs contain only event-driven logs (e.g. incoming and out coming call/SMS), the observed individual sample at a given period represent only a subset of the one observed via signalling traces.

Most importantly, from a temporal perspective, unlike the cross-sectional and non-frequent travel survey data, the main advantages of signalling data are: *i)* the higher temporal granularity and *ii)* the longitudinal nature. Indeed, our methodology presents a proof of concept showing that we can obtain travel survey-level accuracy based on one day signalling data. This incites to go further with the temporal dimension and to apply our approach for extended time periods at the region-scale in order to capture how mobility behaviours evolve over time. Especially, the intra-regional variability of the O-D matrix pattern, such as the weekly or seasonally patterns, are highly relevant for strategic planning whereas very hard and expensive to investigate with traditional travel surveys. Moreover, it is worth to note that the achieved high correlation results of the proposed framework have been obtained with limited-period data (i.e. 24 hours) against several continuous months of data in most existing studies. It follows that our methodology requires much less storage constraints and especially less computational resources (e.g. CPUs, memory, etc.), which represent key aspects for massive data analytics. Hence, that makes our approach more efficient and easier to put into practice for travel demand modelling applications including real-time dynamic O-D estimation taking into account the reduced processing complexity and the long-term applicability of the proposed method.

To the best of our knowledge, the origin-destination matrices generated with signalling data have not been validated at the scale of a territory like Rhône-Alpes region (about 44,000km^2), which contains different socio-demographic and economic territory profiles. Existing works usually focus on cities, small suburban areas or particular transport network roads. Thus, by providing a unique framework valid to a variety of territories (urban and non-urban) and suitable for large-scale mixed environments, our study presents a good trade-off between accuracy and data processing complexity.

Regarding the limitations of our approach, a first challenge to consider, when using mobile phone data, is that they do not include demographic and socio-economic attributes, generally for privacy concerns issues. Although signalling data are characterised by high population coverage, they still lack for complete user information which are required in traditional modelling (e.g. for calibration). Instead, we have shown in this study that the scaling method based solely on census data was sufficient to reveal consistent estimations as the incorporated huge samples may compensate this limitation and the sampling bias in terms of socio-demographic profiles is expected to be marginal with signalling data as they do not strictly depend on the phone usage (as explained earlier). Nevertheless, integrating socio-demographic information within this process would give the opportunity to control the representativeness of different demographics and the result accuracy. In addition, the validation of researches based on cellular data analysis still raises some questions. In our case study, we have used travel survey data as comparable reference to evaluate the empirical analyses. It is our view that, based on the currently available reference data, the methodology outcomes can

only be compared with those of EDR travel survey at an aggregated level. Yet, the two datasets represent different samples and have different attributes. Therefore, additional new methods and tools should be investigated for validation purposes and to better verify the inferred results. In order to give solid interpretations of signalling data-based analyses, it would be relevant to create a set of ground truth data. This could be done by performing a specific controlled experiment (e.g. in some selected zones) aimed to check, with high accuracy and at a disaggregated level, which trips are reported and which ones are missed when using cellular network signalling data.
Furthermore, it shall be noted that the accuracy of the presented approach could be improved by involving longer periods of observations and also by including traces collected from other mobile phone technologies such as the 4G network. That allows to track more detailed movement information about individual trajectories and to support more reliable travel indicators.

7. Conclusions

In this paper, a data-driven modelling approach with novel i) mobile phone massive dataset ii) data pre-processing and iii) validation results is presented. We introduce a full comprehensive workflow of steps to generate origin-destination (O-D) matrices from 2G and 3G cellular network signalling data, which is continuously collected by telecom providers. The proposed method was applied to a dataset of about 2 million cell phone users collected in the Rhône-Alpes region, France. By analysing passive signalling data over a 24-hour period, we show that it is feasible and compelling to use such data in order to estimate O-D matrices that are similar to the ones produced via the travel survey-based method, which provides accurate information about interviewees' trips, but is costlier and time-intensive. An extensive evaluation and validation process with the available travel survey (EDR) data for the Rhône-Alpes region have been performed to deal with the potential of the inferred O-D matrix. Results demonstrate strong similarities with a R^2 coefficient of 0.95 at an aggregated geographical level. This illustrates on one hand the efficiency of our method, as only one day of signalling data has been explored thus reducing considerably its execution time. It shall also be noted that the process can be automated, and then the full workflow could be processed within a few days (e.g., one day) which is significantly faster than conventional travel surveys, leveraging more dynamic applications. On the other hand, our findings show that cell network signalling data can capture unknown and more reasonable flow patterns specifically for low density areas where accurate travel data are either not available or not representative enough. For these reasons, signalling data can be used to support and complement conventional travel surveys as a valuable cost-effective data source for origin-destination estimation, thereby presenting an opportunity to improve significantly and revolutionize the travel demand and traffic flow modelling field.

Signalling data are collected for network management purposes by the providers, thus they are not straightforwardly applicable for mobility and transportation purposes. Pre-processing and filtering of signalling data are essential to make them useful, and this aspect is not very well reported in the literature. In this paper, we propose a pipeline of cell-phone activity indicators-based filtering which ensures a qualitative and large dataset. Indeed, it could be considered as a preliminary guideline to properly process signalling data that could be adapted according to the specific case study.

In addition to the suggested data filtering and processing steps, it will be very interesting to explore how the location accuracy of signalling data, which depends on cell network coverage, can affect different components such as home location and trip detection. Another interesting aspect is to investigate on activity travel patterns such as travel time and how they align with travel survey data. Hence, signalling data temporal resolution needs to be investigated in detail in order to properly define the data potential in terms of spatiotemporal accuracy. As future work, we aim to generate dynamic O-D matrices in order to investigate the travel patterns evolution during different periods of the day based on the same signalling dataset used in this study.

Potential improvements of the proposed workflow will consist in investigating in more detail the assumed hypothesis related to the stationary time threshold and the trip expansion method based on identified home locations from signalling data of a single operator. If an average stationary time could be estimated for each EDR-sector instead of considering one generic threshold, the estimation of movement flows can be further refined. This requires indeed reviewing the trip definition. Although the expansion method gives consistent results, if an accurate penetration rate of 2G and 3G users as well as market share distribution of the data provider within the region are available in the

future, this will allow not only to get more robust estimation of origin-destination matrices but also to infer information at a higher spatial resolution, which can be leveraged for transport planning applications.

Furthermore, with the increasing usage of mobile phones, cell network-based traces are expected to produce even higher-frequency data that will cover a growing number of people, thus allowing for estimating movements at a finer-grained temporal granularity than those provided by travel survey estimations. As a consequence, these emerging individual-based big data could be explored to advance the understanding of less addressed mobility patterns such as during special periods/events. Additionally, signalling data are expected to be available at a national level (e.g.; country level) which depicts a great opportunity to leverage much larger scale patterns. As a result, practical applications relying on recent massive data during long periods and covering unprecedented large areas become feasible.

Exploring cellular network signalling data, together with big data analytics tools for travel demand modelling and transportation planning purposes brings new insights for practitioners, planners and policy-makers to fully benefit from the new promising massive datasets at a low cost, especially with the rising transport networks complexity. They enable to provide prompt response to mobility-related problems and help to keep transport and traffic models updated.

CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

ACKNOWLEDGMENTS

MF's PhD is funded by Orange. The authors would like to thank Orange for giving them access to the mobile phone data collected in the Rhône-Alpes region and Region Auvergne Rhône-Alpes for making the EDR data available. The authors are grateful for the comments by the anonymous reviewers that greatly improved the paper. The authors alone are responsible for the contents of this paper.

AUTHOR CONTRIBUTIONS

MF: Literature search and review, study conception and design, draft manuscript preparation.

MF, TB, ZS, PB, AF and SG: Analysis and interpretation of results, manuscript editing and review.

REFERENCES

Alexander, L., Jiang, S., Murga, M., González, M.C.: Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies. 58, 240–250 (2015). https://doi.org/10.1016/j.trc.2015.02.018

ARCEP: Observatoire des marchés des communications électroniques en France, 4eme trimestre 2017. (2018)

Arentze, T., Timmermans, H.: Data Needs, Data Collection, and Data Quality Requirements of Activity-Based Transport Demand Models. In: Transport Surveys: Raising the Standard. pp. 1–30 (2000)

Asgari, F., Gauthier, V., Becker, M.: A survey on human mobility and its applications. arXiv preprint arXiv:1307.0814. (2013)

Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. EPJ Data Sci. 4, p. 55 (2015). https://doi.org/10.1140/epjds/s13688-015-0046-0

Bohte, W., Maat, K.: Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies. 17, 285–297 (2009). https://doi.org/10.1016/j.trc.2008.11.004

Bonnel, P.: Postal, Telephone, and Face-to-Face Surveys: How Comparable are They? In: Jones, P. and Stopher, P.R. (eds.) Transport Survey Quality and Innovation. pp. 215–237. Emerald Group Publishing Limited (2003)

Bonnel, P.: Prévoir la demande de transport. Presses de l'Ecole Nationale des Ponts et Chaussées, Paris (2004)

Bonnel, P., Fekih, M., Smoreda, Z.: Origin-Destination estimation using mobile network probe data. Transportation Research Procedia. 32, 69–81 (2018). https://doi.org/10.1016/j.trpro.2018.10.013

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., Smoreda, Z.: Passive Mobile Phone Dataset to Construct Origin-Destination Matrix: Potentials and Limitations. Transportation Research Procedia. 11, 381–398 (2015). https://doi.org/10.1016/j.trpro.2015.12.032

Caceres, N., Wideberg, J.P., Benitez, F.G.: Deriving origin–destination data from a mobile phone network. IET Intelligent Transport Systems. 1, 15–26 (2007). https://doi.org/10.1049/iet-its:20060020

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C.: Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. IEEE Transactions on Intelligent Transportation Systems. 12, 141–151 (2011)(a). https://doi.org/10.1109/TITS.2010.2074196

Calabrese, F., Di Lorenzo, G., Liu, L., Ratti, C.: Estimating Origin-Destination Flows Using Mobile Phone Location Data. IEEE Pervasive Computing. 10, 36–44 (2011)(b). https://doi.org/10.1109/MPRV.2011.41

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C.: Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies. 26, 301–313 (2013)

Calabrese, F., Ferrari, L., Blondel, V.D.: Urban Sensing Using Mobile Phone Network Data: A Survey of Research. ACM Computing Surveys. 47, 1–20 (2014). https://doi.org/10.1145/2655691

CERTU: L'enquête ménages déplacements standard CERTU, éditions du CERTU. (2008)

Chen, C., Bian, L., Mac, J.: From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transportation Research Part C: Emerging Technologies. 46, 326–337 (2014). https://doi.org/10.1016/j.trc.2014.07.001

Chen, C., Gong, H., Lawson, C., Bialostozky, E.: Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A: Policy and Practice. 44, 830–840 (2010). https://doi.org/10.1016/j.tra.2010.08.004

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M.: The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies. 68, 285–299 (2016). https://doi.org/10.1016/j.trc.2016.04.005

Choujaa, D.: Activity Recognition from Mobile Phone Data: State of the Art, Prospects and Open Problems. In: Imperial college London. p. 32 (2009)

Deutsch, K., McKenzie, G., Janowicz, K., Li, W., Hu, Y., Goulias, K.: Examining the use of smartphones for travel behavior data collection. Presented at the Conference of the International Association for Travel Behavior Research, Torento, Canada (2012)

Feng, T., Timmermans, H.J.P.: Extracting Activity-travel Diaries from GPS Data: Towards Integrated Semi-automatic Imputation. Procedia Environmental Sciences. 22, 178–185 (2014). https://doi.org/10.1016/j.proenv.2014.11.018

Fiadino, P., Ponce-Lopez, V., Antonio, J., Torrent-Moreno, M., D'Alconzo, A.: Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era." In: Big-DAMA. pp. 43–48. ACM Press (2017)

Fiadino, P., Valerio, D., Ricciato, F., Hummel, K.A.: Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data. In: Pescapè, A., Salgarelli, L., and Dimitropoulos, X. (eds.) Traffic Monitoring and Analysis. pp. 66–80. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)

Frias-Martinez, V., Virseda, J., Rubio, A., Frias-Martinez, E.: Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. In: Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development. p. 11. ACM (2010)

Giannotti, F., Pedreschi, D. eds: Mobility, data mining and privacy: geographic knowledge discovery. Springer, Berlin (2008)

González, M.C., Hidalgo, C.A., Barabási, A.-L.: Understanding individual human mobility patterns. Nature. 453, 779–782 (2008). https://doi.org/10.1038/nature06958

Gonzalez, P.A., Weinstein, J.S., Barbeau, S.J., Labrador, M.A., Winters, P.L., Georggi, N.L., Perez, R.: Automating mode detection for travel behaviour analysis by using global positioning systems-enabled mobile phones and neural networks. IET Intelligent Transport Systems. 4, 37–49 (2010)

Graells-Garrido, E., Saez-Trumper, D.: A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow. In: 2nd International conference on IoT in Urban Space. pp. 1–7. ACM Press (2016)

Gundlegård, D., Rydergren, C., Breyer, N., Rajna, B.: Travel demand estimation and network assignment based on cellular network data. Computer Communications. 95, 29–42 (2016). https://doi.org/10.1016/j.comcom.2016.04.015

Hoteit, S., Chen, G., Viana, A.C., Fiore, M.: Filling the Gaps: On the Completion of Sparse Call Detail Records for Mobility Analysis. In: Eleventh ACM Workshop on Challenged Networks (2016)

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., Pujolle, G.: Estimating human trajectories and hotspots through mobile phone data. Computer Networks. 64, 296–307 (2014)

Huang, L., Li, Q., Yue, Y.: Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks - LBSN '10. p. 27. ACM Press, San Jose, California (2010)

Huang, Z., Ling, X., Wang, P., Zhang, F., Mao, Y., Lin, T., Wang, F.-Y.: Modeling real-time human mobility based on mobile phone and transportation data fusion. Transportation Research Part C: Emerging Technologies. 96, 251–269 (2018). https://doi.org/10.1016/j.trc.2018.09.016

Iqbal, Md.S., Choudhury, C.F., Wang, P., González, M.C.: Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies. 40, 63–74 (2014). https://doi.org/10.1016/j.trc.2014.01.002

Janzen, M., Vanhoof, M., Smoreda, Z., Axhausen, K.W.: Closer to the total? Long-distance travel of French mobile phone users. Travel Behaviour and Society. 11, 31–42 (2018). https://doi.org/10.1016/j.tbs.2017.12.001

Jiang, S., Ferreira, J., González, M.C.: Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. IEEE Transactions on Big Data. 3, 208–219 (2017)

Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr, J., Frazzoli, E., González, M.C.: A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing. p. 9. ACM (2013)

Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., González, M.C.: The TimeGeo modeling framework for urban mobility without travel surveys. Proceedings of the National Academy of Sciences. 113, E5370–E5378 (2016). https://doi.org/10.1073/pnas.1524261113

Mellegard, E., Moritz, S., Zahoor, M.: Origin/Destination-estimation Using Cellular Network Data. In: 11th International Conference on Data Mining Workshops (ICDMW). pp. 891–896. IEEE (2011)

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

Naboulsi, D., Fiore, M., Ribot, S., Stanica, R.: Large-Scale Mobile Traffic Analysis: A Survey. IEEE Communications Surveys Tutorials. 18, 124–161 (2016). https://doi.org/10.1109/COMST.2015.2491361

Ni, L., Wang, X. (Cara), Chen, X. (Michael): A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. Transportation Research Part C: Emerging Technologies. 86, 510–526 (2018). https://doi.org/10.1016/j.trc.2017.12.002

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P.: Supporting large-scale travel surveys with smartphones – A practical approach. Transportation Research Part C: Emerging Technologies. 43, 212–221 (2014). https://doi.org/10.1016/j.trc.2013.11.005

Nour, A., Hellinga, B., Casello, J.: Classification of automobile and transit trips from Smartphone data: Enhancing accuracy using spatial statistics and GIS. Journal of Transport Geography. 51, 36–44 (2016). https://doi.org/10.1016/j.jtrangeo.2015.11.005

Ortúzar, J. de D., Willumsen, L.G.: Modelling Transport Juan de Dios Ortúzar, Luis G. Willumsen. John Wiley & Sons (2011)

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Using Mobile Phones to Determine Transportation Modes. ACM Trans. Sen. Netw. 6, 13:1–13:27 (2010). https://doi.org/10.1145/1689239.1689243

Ricciato, F., Widhalm, P., Pantisano, F., Craglia, M.: Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. Pervasive and Mobile Computing. 35, 65–82 (2016). https://doi.org/10.1016/j.pmcj.2016.04.009

Schlaich, J., Otterstätter, T., Friedrich, M.: Generating Trajectories from Mobile Phone Data. In: TRB 89th Annual Meeting Compendium of Papers. p. 18. , Washington, D.C., USA (2010)

Shen, L., Stopher, P.R.: Review of GPS Travel Survey and GPS Data-Processing Methods. Transport Reviews. 34, 316–334 (2014). https://doi.org/10.1080/01441647.2014.903530

Smoreda, Z., Olteanu-Raimond, A.-M., Couronné, T.: Spatiotemporal data from mobile phones for personal mobility assessment. In: Transport survey methods: best practice for decision making. pp. 745–768. Emerald Group Publishing Limited (2013)

Stopher, P., FitzGerald, C., Zhang, J.: Search for a global positioning system device to measure person travel. Transportation Research Part C: Emerging Technologies. 16, 350–369 (2008). https://doi.org/10.1016/j.trc.2007.10.002

Stopher, P.R., Greaves, S.P.: Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice. 41, 367–381 (2007). https://doi.org/10.1016/j.tra.2006.09.005

Tettamanti, T., Demeter, H., Varga, I.: Route Choice Estimation Based on Cellular Signaling Data. Acta Polytechnica Hungarica. 9, 207–220 (2012)

Tettamanti, T., Istvan, V.: Mobile phone Location Area Based Traffic Flow estimation in urban road traffic. Advances in Civil and Environment Engineering. 1, 1–15 (2014)

Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C.: The path most traveled: Travel demand estimation using big data resources. Transportation Research Part C: Emerging Technologies. 58, 162–177 (2015). https://doi.org/10.1016/j.trc.2015.04.022

Wang, F., Chen, C.: On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies. 87, 58–74 (2018). https://doi.org/10.1016/j.trc.2017.12.003

Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C.: Understanding Road Usage Patterns in Urban Areas. Scientific Reports. 2, 1–6 (2012). https://doi.org/10.1038/srep01001

Wang, Z., He, S.Y., Leung, Y.: Applying mobile phone data to travel behaviour research: A literature review. Travel Behaviour and Society. (2017). https://doi.org/10.1016/j.tbs.2017.02.005

White, J., Wells, I.: Extracting origin destination information from mobile phone data. In: Eleventh International Conference on Road Transport Information and Control, 2002. (Conf. Publ. No. 486). pp. 30–34 (2002)

Widhalm, P., Nitsche, P., Brändie, N.: Transport mode detection with realistic Smartphone sensor data. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 573–576 (2012)

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M.C.: Discovering urban activity patterns in cell phone data. Transportation. 42, 597–623 (2015). https://doi.org/10.1007/s11116-015-9598-x

Wismans, L.J.J., Friso, K., Rijsdijk, J., de Graaf, S.W., Keij, J.: Improving A Priori Demand Estimates Transport Models using Mobile Phone Data: A Rotterdam-Region Case. Journal of Urban Technology. 25, 63–83 (2018). https://doi.org/10.1080/10630732.2018.1442075

Wolf, J., Oliveira, M., Thompson, M.: Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey. Transportation Research Record: Journal of the Transportation Research Board. 1854, 189–198 (2003). https://doi.org/10.3141/1854-21

Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., Axhausen, K.W.: Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information. Transportation Research Record: Journal of the Transportation Research Board. 1870, 46–54 (2004). https://doi.org/10.3141/1870-06

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., Li, Q.: Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation. 42, 625–646 (2015). https://doi.org/10.1007/s11116-015-9597-y

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., Yin, L.: Understanding the bias of call detail records in human mobility research. International Journal of Geographical Information Science. 30, 1738–1762 (2016). https://doi.org/10.1080/13658816.2015.1137298